Copyright

by

Fanglei Zhuang

2009

**The Dissertation Committee for Fanglei Zhuang Certifies that this is the approved version of the following dissertation:**

# Group II Intron Mobility and its Gene Targeting Applications in Prokaryotes and Eukaryotes

**Committee:**

Alan M. Lambowitz, Supervisor

Scott W Stevens

John B Wallingford

Philip W Tucker

Makkuni Jayaram

# Group II Intron Mobility and its Gene Targeting Applications in Prokaryotes and Eukaryotes

**by**

**Fanglei Zhuang, B.S.**

**Dissertation**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**Doctor of Philosophy**

**The University of Texas at Austin**

**August 2009**

# Dedication

To my parents and husband for their support and encouragement.

## Acknowledgements

I owe my gratitude to all those people who have made this dissertation possible.

This is a great opportunity to express my gratitude to my advisor, Dr. Alan M. Lambowitz, for his support, great academic guidance and patience. His diligence and persistence in science will influence my whole life.

I would like to thank Dr. John B Wallingford, who has been always there to offer help and advice.

I am also very grateful for having an exceptional doctoral committee and want to thank Dr. Scott W Stevens, Dr. Makkuni Jayaram, Dr. John B Wallingford, and Dr. Philip W Tucker for their guidance and encouragement.

I thank all members of Lambowitz lab for their love and help. Special thanks go to Marta Mastroianni. It was such a great experience to work with her.

I am very pleased to know everyone in Wallingford lab too, who taught me to use their instruments and take care of frogs.

Most importantly, none of this work would have been possible without my husband, Jian Song. His continuous support, love and patience help me through the difficulties. His love keeps my life in proper perspective and balance.

# Group II Intron Mobility and its Gene Targeting Applications in Prokaryotes and Eukaryotes

Publication No._____

Fanglei Zhuang

The University of Texas at Austin, 2009

Supervisor: Alan M. Lambowitz

Mobile group II introns are retroelements that insert site-specifically into DNA target sites by a process called retrohoming. Retrohoming is mediated by a ribonucleoprotein particle (RNP) that contains both the intron RNA and the intron-encoded protein (IEP). My dissertation focuses on two mobile group II introns: *Lactococcus lactis* Ll.LtrB and *Escherichia coli* EcI5, which belong to structural subclasses IIA and CL/IIB1, respectively. Previous studies showed that the Ll.LtrB IEP, denoted LtrA protein, is pole localized in *E. coli*. First, I found that active LtrA protein is associated with *E. coli* membrane fractions, suggesting that LtrA pole localization might reflect association with a membrane receptor. Second, I found that EcI5 is highly active in retrohoming in *E. coli* and obtained a comprehensive view of its DNA target site recognition by selection experiments. I found that EcI5 recognizes DNA target sequences by using both the IEP and base pairing of the intron RNA, with the IEP having different target specificity than for other mobile group II introns. A computer algorithm based on the empirically determined DNA recognition rules enabled retargeting of EcI5 to

integrate at ten different sites in the chromosomal *lacZ* gene at frequencies up to 98% without selection. Finally, I developed methods for gene targeting in the frog *Xenopus laevis* by using Ll.LtrB RNPs for site-specific DNA modification in isolated sperm nuclei, followed by *in vitro* fertilization to generate genetically modified animals. The site-specific integrations were efficient enough to detect in fifty sperm nuclei for a multiple copy target site, the Tx1 transposon, and several hundred sperm nuclei for protein-encoding genes. Based on these results, I obtained transgenic tadpoles with site-specific Tx1 integrations by simple screening. To facilitate screening for embryos with targeted integrations in protein-encoding genes, I constructed an intron carrying a GFP-RAM (Retrotransposition-Activated Marker). By using this GFP-RAM with introns containing randomized sequences that base pair with the target DNA, I obtained tadpoles with intron integrations at different genomic locations, including protein-encoding genes. The methods for using group II introns for targeted sperm DNA modification in *X. laevis* may be applicable to other animals.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1: Introduction

## 1.1 GROUP II INTRONS

Group II introns are mobile elements found in prokaryotes and eukaryotic organelles (Pyle and Lambowitz 2006). They are thought to be ancestors of spliceosomal introns and non-LTR retrotransposons in higher organisms. Group II introns consist of an autocatalytic intron RNA and a multi-functional intron-encoded protein (IEP), which has reverse transcriptase (RT) activity. The IEP acts together with the intron RNA to promote RNA splicing and reverse splicing into double-stranded DNA (retrohoming) by stabilizing the catalytically active RNA structure (Lambowitz and Zimmerly 2004; Pyle and Lambowitz 2006). Thus far, more than 200 group II introns have been identified, with the number increasing as new genomes are sequenced (Dai et al. 2003; Simon et al. 2008).

### 1.1.1 Group II intron RNA structure

The conserved group II intron RNA secondary structure contains six double-helical domains (Domain I-VI) extending from a central wheel (Michel and Ferat 1995) (see Figure 1.1). Domain I (DI), an essential domain, is involved in catalysis and the recognition of exons through base-pairing interactions (Mills et al. 1996; Mohr et al. 2000; Pyle and Lambowitz 2006). These interactions involve intron RNA sequences denoted exon-binding sites 1 and 2 (EBS1 and 2) within DI, which bases pair with 5'-exon sequences denoted intron binding sites 1 and 2 (IBS1 and 2) (Michel and Ferat 1995). In group IIA introns, a third sequence denoted δ in DI bases pairs with the 3'-exon sequence denoted δ', while in group IIB and IIC introns, a different DI sequence denoted EBS3 base pairs with the 3'-exon sequence, now denoted IBS3 (Michel and Ferat 1995;

Costa et al. 2000; Lambowitz and Zimmerly 2004). These base-pairing interactions position the 5'- and 3'-exon sequences at the intron's active site and are required both for intron forward splicing and reverse splicing. Domain II (DII) contributes to intron RNA folding by providing essential tertiary contacts with DI (Chanfreau and Jacquier 1996; Costa et al. 1997; Fedorova et al. 2003). Domain III (DIII) is referred to as a catalytic effector (Pyle and Lambowitz 2006). It is not strictly required for catalysis (Koch et al. 1992), but its presence substantially enhances the reaction rates of group II intron-derived ribozyme constructs (Fedorova et al. 2003). Among group II introns, the open reading frame (ORF) of the IEP is invariably located in Domain IV (DIV), outside the intron's catalytic core. The IEP assists intron splicing under physiological conditions and is required for both intron splicing and retrohoming (Lambowitz and Zimmerly 2004). The subdomain DIVa is the high-affinity binding site for the IEP (Wank et al. 1999). Domain V (DV), another essential domain for intron catalysis, is the most conserved substructure of the intron (Michel and Ferat 1995). Sequences in DV interact with other domains to facilitate *trans*-splicing and coordinate divalent metal ions for catalysis (Jarrell et al. 1988; Toor et al. 2008). Domain VI (DVI) contains a bulged adenosine residue that serves as the branch point during intron forward splicing (Michel and Ferat 1995; Lambowitz and Zimmerly 2004; Pyle and Lambowitz 2006).

### 1.1.2 Group II intron lineages

Based on RNA structure and sequence characteristics, three major classes of group II intron RNAs, denoted IIA, IIB, and IIC, have been distinguished (Figure 1.1) (Michel et al. 1989; Dai and Zimmerly 2002a; Toro 2003; Toro et al. 2007; Simon et al. 2008). The major differences between IIA and IIB introns are indicated in the boxed region of Figure 1.1. The differences include the ε' region in the internal loop C in DI,

the location of EBS2, and the presence of EBS3. EBS2 is located within a stem-loop in group IIA introns, in a bend or bulge within group IIB introns (Toor et al. 2001), and is absent in group IIC introns. Additionally, the 5'-exons of many group IIC introns resemble transcription terminators, a feature not present in group IIA and IIB introns (Granlund et al. 2001; Dai and Zimmerly 2002a). This may reflect an additional interaction required for splicing or mobility that is specific to group IIC introns. The EBS3 sequence is present in both group IIB and IIC introns, but absent in IIA introns. In group IIB and IIC introns, this sequence consists of one nucleotide residue that resides within a bulge in DI and interacts with the IBS3 in the 3'-exon. In group IIA introns, the δ sequence is adjacent to IBS1 and it base pairs with the δ' sequence in 3'-exon.

Mobile group II intron RNAs have co-evolved with their respective IEPs. Based on phylogenetic analysis of intron IEPs, group II introns are classified by lineages denoted chloroplast-like (CL), mitochondrial-like (ML), and bacterial A-F (Toor et al. 2001; Zimmerly et al. 2001). CL and ML group II introns are found both in prokaryotes and in eukaryotic mitochondria and chloroplasts, possibly reflecting their association with bacterial endosymbionts that gave rise to these organelles, while bacterial A-F introns are found only in eubacteria and archaea (Toor et al. 2001).

### 1.1.3 Group II intron splicing reaction

Group II introns are ribozymes that catalyze their own splicing. Similarities in the splicing reaction mechanisms, splice site consensus sequences, and the structure and function of RNA domains provide evidence for an evolutionary link between spliceosomal introns and group II introns (Matsuura et al. 2001).

Most group II introns splice via RNA-catalyzed transesterification reactions involving formation of a lariat intermediate. Some group II introns can self-splice *in vitro*

under non-physiological conditions with high salt, high $Mg^{2+}$ concentrations and high temperature (Schmelzer and Schweyen 1986; Tabak et al. 1987; Perlman and Podar 1996). In the branching pathway, the splicing mechanism consists of two consecutive transesterification reactions (Figure 1.2). In the first reaction, the 2'-hydroxyl of a bulged adenosine in the DVI stem attacks the 5'-splice site, creating a free 5'-exon and a branched intermediate consisting of intron lariat with the 3'-exon. In the second reaction, the 5'-exon attacks the 3'-splice site, leading to the ligation of the exons and concomitant release of the lariat intron. The excised intron lariat has a 2'-5' linkage and a short tail (Lambowitz and Zimmerly 2004).

Some group II introns can also self-splice *in vitro* through another pathway called hydrolytic splicing. In this pathway, a water molecule acts as the nucleophile in the first step and attacks the phosphate at the 5'-splice site. The resulting intermediate is a linear intron with the 3'-exon. The second step of splicing is the same as in the branching pathway, but now generates a linear intron and ligated exons (Daniels et al. 1996).

### 1.1.4 Group II intron-encoded proteins (IEPs)

About one-third of the known group II introns encode a multi-functional IEP with homology to reverse transcriptase (Lambowitz and Zimmerly 2004). The IEP functions in RNA splicing by stabilizing the catalytically active RNA structure and then remains bound to the excised intron RNA in an RNP complex. The RNA and IEP function together to carry out reverse splicing into DNA target sites during intron mobility (Guo et al. 2000; Matsuura et al. 2001; Noah et al. 2006).

The IEPs are structurally conserved in evolution (Michel and Ferat 1995). The most detailed biochemical analysis has been done on the IEPs of group IIA introns, specifically the yeast mitochondrial introns *cox1*-I1 and -I2 and the *Lactococcus lactis*

LtrB (Ll.LtrB) intron. The IEPs of these introns contain four domains denoted reverse transcriptase (RT), a domain required for RNA splicing (maturase) activity (X), DNA binding (D), and DNA endonuclease (En) (Kennell et al. 1993; Mohr et al. 1993; Matsuura et al. 1997; Saldanha et al. 1999; San Filippo and Lambowitz 2002). The N-terminal RT domain contains the conserved sequence regions, RT-1 to -7, which are found in retroviral and other RTs. RT-0 is at the N-terminus of the RT domain and is characteristic of the RTs of non-LTR retrotransposons (Malik et al. 1999; Zimmerly et al. 1999). Domain X is a site of mutations that affect maturase activity in which the IEP binds specifically to the intron RNA and stabilizes the catalytically active RNA structure. It functions together with the RT domain to bind the intron RNA during RNA splicing and reverse splicing (Cui et al. 2004; Lambowitz and Zimmerly 2004; Blocker et al. 2005). The D domain is critical for DNA target site binding in reverse splicing. The deletion of this region strongly hinders intron RNA reverse splicing into double-stranded DNA (San Filippo and Lambowitz 2002). The En domain is related to bacterial colicin and pyocin nucleases (Gorbalenya 1994). It is capable of cleaving the bottom strand of the double stranded DNA target site to generate a primer for reverse transcription (Guo et al. 1997; Singh and Lambowitz 2001).

Notably, many bacterial group II introns encode IEPs lacking the En domain. Those En-deficient introns are still able to retrohome by reverse splicing into double- or single-stranded DNA and utilizing nascent strands at DNA replication forks to prime reverse transcription (Martinez-Abarca and Toro 2000; Dai and Zimmerly 2002a; Zhong and Lambowitz 2003).

## 1.2 *LACTOCOCCUS LACTIS* LL.LTRB INTRON

The *L. lactis* Ll.LtrB group IIA intron has been studied extensively as a model system for understanding group II intron mobility mechanisms (Figure 1.5). This intron was discovered in a relaxase gene (*ltrB*) in an *L. lactis* conjugative element pRS01, and its splicing is essential to generate functional relaxase for conjugation (Mills et al. 1996; Shearman et al. 1996).

Ll.LtrB encodes a multiple functional IEP, denoted LtrA protein, which has been a model system for group II IEPs. It contains four conserved domains: RT, X, D, and En. LtrA binds specifically to the Ll.LtrB intron RNA and promotes RNA splicing by promoting the formation of critical RNA tertiary interactions (Matsuura et al. 2001). It also plays a crucial role in retrohoming through target DNA-primed reverse transcription (Singh and Lambowitz 2001).

The retrohoming reaction of Ll.LtrB is carried out by a RNP complex that is formed during RNA splicing and is composed of LtrA and spliced intron lariat RNA (Singh and Lambowitz 2001; Lambowitz and Zimmerly 2004; Noah et al. 2006). To initiate the reverse splicing, Ll.LtrB RNPs bind DNA and recognize target sites (Figure 1.3) (Aizawa et al. 2003). The initial recognition is thought to occur through major groove interactions of LtrA with a small number of specific bases in the distal 5'-exon region, including T-23, G-21, and A-20 (Figure 1.4). These base interactions, together with neighboring phosphate-backbone contacts and possibly minor groove interactions with positions -18 and -14, result in local DNA melting, which enables the intron RNA to base pair with a 13-15 nt region of the DNA target site encompassing the intron-insertion site. The intron RNA sequences involved in this base pairing are EBS1, 2, and δ, and the complementary DNA target-site sequences are IBS1 and 2 in the 5'-exon and δ' in the 3'-

exon. After binding and base paring, the intron RNA reverse splices itself into the intron-insertion site at the junction of the 5'- and 3'-exon of the DNA target site. LtrA then cleaves the bottom strand between position +9 and +10 to generate the primer for reverse transcription. This second-strand cleavage requires additional interactions between LtrA and the 3'-exon, of which the most critical is recognition of +5T position in the 3'-exon (Mohr et al. 2000; Singh and Lambowitz 2001). After second-strand cleavage, LtrA uses the 3' end at the cleavage site as a primer for reverse transcription and synthesizes a full-length intron cDNA, which is fully integrated into the DNA target site by host DNA repair mechanisms (Mills et al. 1997; Cousineau et al. 1998; Smith et al. 2005).

In the absence of second-strand cleavage, the Ll.LtrB intron can still retrohome at a low frequency by using nascent leading or lagging strands at DNA replication forks as primers for reverse transcription (Zhong and Lambowitz 2003). A similar mechanism is also used for retrohoming by group II introns that encode IEPs lacking the En domain (referred as $En^-$ introns) (Lambowitz and Zimmerly 2004). Most $En^-$ introns are found on the lagging template strand and appear to preferentially utilize nascent lagging strands as primers for cDNA synthesis (Martinez-Abarca et al. 2004).

Recent studies on LtrA localization in *E. coli* and *L. lactis* revealed that LtrA localizes to the cellular poles with or without co-expression of the Ll.LtrB intron RNA (Zhao and Lambowitz 2005). This polar localization of LtrA in *E. coli* could account for the preferential insertion of Ll.LtrB introns with randomized EBS and δ sequences into the chromosomal *ori* and *ter* regions, which are also pole localized during much of the cell cycles (Niki and Hiraga 1998; Draper and Gober 2002; Zhong et al. 2003; Zhao and Lambowitz 2005). Recent studies using an LtrA-GFP fusion and cellular microarrays to screen an *E. coli* transposon insertion library for mutations that alter LtrA localization

patterns showed that polyphosphate accumulation can affect the polar localization of LtrA and other basic proteins in bacteria (Zhao et al. 2008).

### 1.3 *ESCHERICHIA COLI* EcI5 INTRON

The *E. coli* EcI5 intron is a group IIB intron. It was discovered in a virulence plasmid pO157 in the pathogenic *E. coli* strain O157:H7 (Burland et al. 1998). The intron is inserted within a non-coding region of the plasmid, but in the same orientation as most other plasmid genes. Southern hybridizations and PCR analysis showed that EcI5 is also inserted at the same site in DNA isolated from 19 of 72 ECOR strains, although always in fragmented form (Dai and Zimmerly 2002b).

The EcI5 intron has a standard IIB intron structure (Figure 1.6A), as previously discussed (see section 1.1). It contains EBS 1, 2 and 3 in DI, which pair with IBS1 and 2 in the 5'-exon and IBS3 in the 3'-exon. These base-pairing interactions are required for intron splicing and the target site recognition during intron retrohoming (Zhuang et al. 2009). The EcI5 IEP has all four conserved domains, RT, X, D, and En, shared with other group II IEPs (Figure 1.6B). Notably, the EcI5 intron is the only *E. coli* intron whose IEP encodes an En domain, which can potentially cleave the opposite strand of the double-stranded DNA target site to generate a primer for reverse transcription (Zimmerly et al. 2001).

### 1.4 GROUP II INTRON BASED "TARGETRON" TECHNOLOGY

The retrohoming process of group II introns is mediated by the RNP complex containing the IEP and the spliced intron RNA. It is known that both IEP and RNA are involved in the DNA target site recognition (Guo et al. 1997; Matsuura et al. 2001; Singh and Lambowitz 2001; Noah et al. 2006). Because base pairing of the intron RNA to the DNA target site contributes most of the DNA target specificity, one can retarget group II

introns to the desired sites by modifying the exon-recognition sequences in the intron RNA (Lambowitz et al. 2005). Our laboratory used the Ll.LtrB intron to develop "targetron" technology, which has been widely used for targeted gene disruption and site-specific DNA insertion in diverse Gram-negative and Gram-positive bacteria (Karberg et al. 2001; Frazier et al. 2003; Zhong et al. 2003; Perutka et al. 2004; Chen et al. 2005; Yao et al. 2006; Pearson and Mobley 2007; Shao et al. 2007; Yao and Lambowitz 2007; Malhotra and Srivastava 2008).

For gene targeting in bacteria, the Ll.LtrB intron is introduced via a donor plasmid, which expresses an Ll.LtrB-ΔORF intron with short flanking exons. The Ll.LtrB-ΔORF intron has most of the LtrA ORF deleted from DIV, and the LtrA protein is expressed instead from a position just downstream of the 3'-exon (Guo et al. 2000; Karberg et al. 2001). Based on previous genetic selection experiments, a computer algorithm was developed to scan the DNA target sequence for the best matches to the positions recognized by the IEP. PCR primers are then designed to modify the intron's EBS and δ regions to base pair optimally to the IBS and δ' region in the target site (Perutka et al. 2004). The positions recognized by the IEP are sufficiently few and flexible that the algorithm readily identifies multiple rank-ordered target sites in any gene. Furthermore, the intron can be targeted to insert into different orientations by targeting the sense or anti-sense strand of DNA target sites, making it possible to obtain either unconditional or conditional disruptions (Frazier et al. 2003; Yao et al. 2006). Selectable markers can be incorporated into the intron DIV in the place of the IEP ORF to facilitate the identification of desired integration events, which can be detected with nearly 100% efficiently after selection (Zhong et al. 2003). Additionally, targetrons with

randomized exon-recognition sequences can be used to generate libraries of random disruptants, analogous to transposon mutagenesis (Zhong et al. 2003).

In addition to bacteria, "targetron" technology can potentially be applied in eukaryotes. The Ll.LtrB intron was retargeted to the HIV-1 provirus and the human gene encoding the HIV-1 co-receptor CCR5 with high frequency and specificity in *E. coli* plasmid assays (Guo et al. 2000). These retargeted group II intron RNPs retained activity in human cell assays, in which plasmid-based target sites and group II intron RNPs were introduced separately by liposome-mediated transfection. In these experiments, however, the group II intron integration reaction in human cells was much less efficient than in bacteria, requiring nested PCR for detection (Guo et al. 2000). *E. coli* assays also show that group II introns can be used for genetic repair, either by inserting a functional copy of the gene into a defective gene, or by inserting functional exons preceded by a splice acceptor site to avoid the defective exons located downstream (Jones et al. 2005).

A recent *Xenopus laevis* oocyte microinjection assay showed that group II intron RNPs can insert efficiently into a plasmid target site in oocyte nuclei, but only after injection of additional $Mg^{2+}$ (Mastroianni et al., 2008). In such assays under optimal conditions group II intron RNP integration frequencies were as high as 38% and targeting frequencies by double-strand break stimulated homologous recombination were as high as 4.8%. Interestingly, group II intron RNPs reconstituted *in vitro* with linear as well as lariat intron RNA can be used to introduce the double-strand break (Mastorianni et al. 2008). Similar RNP microinjection assays showed relatively efficient $Mg^{2+}$-dependent group II intron insertion into plasmid target sites in *D. melanogaster* and zebrafish embryos, as well as integration into chromosomal target sites in *D. melanogaster*

embryos (Mastroianni et al. 2008). These results demonstrate the potential use of group II introns for gene targeting in higher organisms.

### 1.5 OVERVIEW OF DISSERTATION RESEARCH

This dissertation research is focused on two group II introns: the group IIA intron Ll.LtrB from *Lactococcus lactis* and the group IIB intron EcI5 from *E. coli*. The dissertation is divided into three related subjects. First, following up the previous finding of polar localization of LtrA in *E. coli*, which raised the possibility of a membrane-attachment site, I analyzed the distribution of the LtrA protein in *E. coli* subcellular fractions by using bacterial cell fractionation methods. These experiments showed that the majority of LtrA protein is found in membrane fractions. Further assays using sucrose gradient equilibrium centrifugation showed that LtrA can be found in both inner and outer membrane fractions. I also showed that the LtrA protein in membrane fractions is active in promoting the splicing of its own RNA precursor. Second, I analyzed the DNA target site sequences required for retrohoming of the EcI5 intron, including the sequences recognized by the IEP and base pairing of the intron RNA. Based on these target site recognition rules, I demonstrated that the EcI5 intron could be used for gene targeting in bacteria and appears to be more efficient and have higher specificity than does the Ll.LtrB intron. Finally, I used the Ll.LtrB intron for site-specific modification of genomic DNA in decondensed *X. laevis* sperm nuclei and then used the modified sperm nuclei for *in vitro* fertilization to generate genetically modified animals. I first optimized the RNP reaction conditions for efficient targeting in *X. laevis* sperm nuclei and developed conditions under which the modified sperm nuclei were able to fertilize eggs to yield viable embryos and tadpoles. Further, by using the Ll.LtrB intron with randomized EBS and δ sequence and a GFP marker inserted in intron DIV, I showed the Ll.LtrB intron

could be used not only for site-specific gene targeting but also in forward genetic screens to study gene function in *X. laevis*. This work produced the first viable eukaryotic organisms generated by group II intron gene targeting.

This dissertation is written in four chapters. In the next chapter (Chapter 2), the LtrA protein distribution profile in *E. coli* subcellular fractions will be described. I found that 83% of LtrA is associated with cellular membrane fractions. By separating the inner and outer membrane, I found that 30% co-purified with the outer membrane and only 6% co-purified with the inner membrane. Further, biochemical assays showed that LtrA protein in membrane fractions is capable of promoting the *in vitro* splicing of Ll.LtrB RNA precursor at relatively low salt conditions *in vitro*. In Chapter 3, I will describe my research on DNA target site recognition by the EcI5 intron. In these experiments, I found that the DNA target site for the EcI5 intron extends from position –26 to +10 from the intron insertion site, with a 12-nt region from –13 to 9 and -6 to +1 recognized by base pairing of the intron RNA. The IEP recognizes a small number of bases flanking those recognized by the intron RNA, but their identity is different from those in previously characterized group II introns. A computer algorithm based on the empirically determined DNA target site recognition rules enabled retargeting of EcI5 to integrate efficiently and specifically at different sites in the *E. coli* chromosomal *lacZ* gene. Chapter 4 describes the development of group II intron-based gene targeting methods for *X. laevis*. I show that the Ll.LtrB intron can integrate site-specifically into the desired target sites in the sperm genome (Tx1, mitf and tyr). The targeting reaction in sperm nuclei is dependent on the temperature, reaction time and the decondensation level of sperm chromatin by nucleoplasmin. The modified sperm nuclei were then used for *in vitro* fertilization to generate transgenic animals. To facilitate screening larger numbers of

embryos for targeting protein-encoding genes, a GFP-RAM (Retrotransposition-Activated Marker), which is expressed only after DNA integration, was incorporated into intron DIV. The GFP-RAM intron construct allowed me to perform a screen of efficient targetrons in *X. laevis* using an intron library with randomized EBS and δ sequences. My results also suggest that this method using an intron library and GFP-RAM marker could be employed for forward genetic screens for analysis of gene function in *X. laevis*. The technology I developed for group II intron-based gene targeting by sperm DNA modification in *X. laevis* has the potential to be broadly applicable to other animal species, including zebrafish, mice, livestock, and primates.

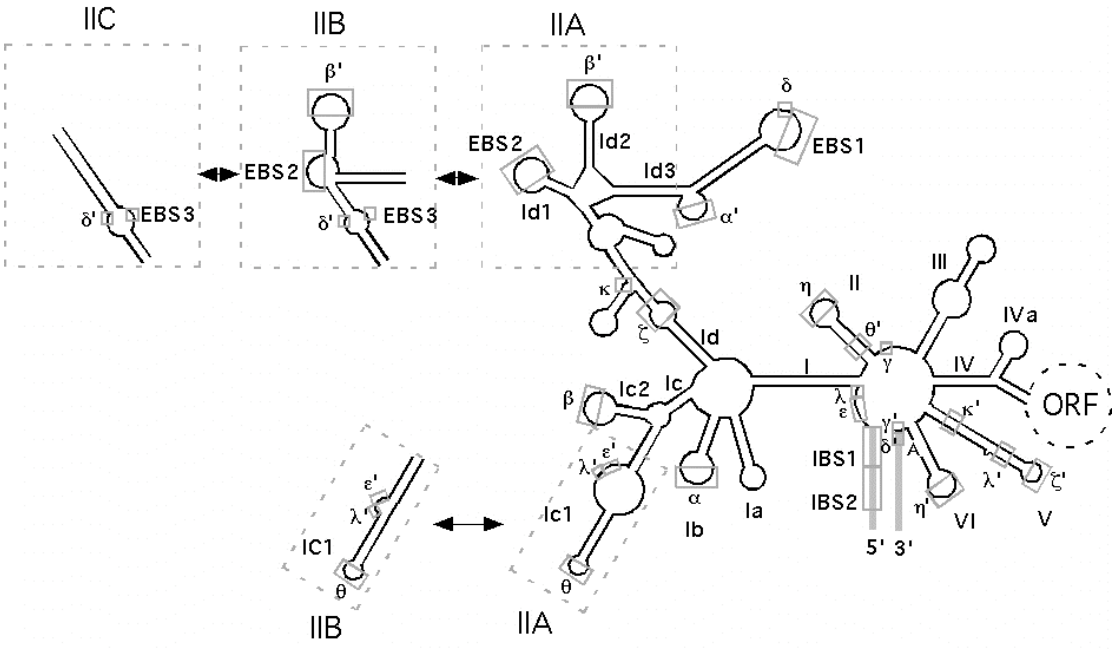



Figure 1.1: Group II intron RNA secondary structure.

The conserved secondary structure of group II introns consists of six double-helical domains (DI-DVI) emanating from a central wheel, with sub-domains indicated by lower-case letter (e.g., DIa, DIVa). The ORF is encoded within DIV (dotted loop), and DIVa is the high-affinity binding site for the intron-encoded protein (IEP). Greek letters indicate sequences involved in tertiary interactions. EBS and IBS refer to exon- and intron-binding sites, respectively. The sequences involved in tertiary interactions are boxed. Some key differences between subgroup IIA, IIB and IIC introns are indicated within dashed boxes. The figure was adapted from Lambowitz and Zimmerly (2004).

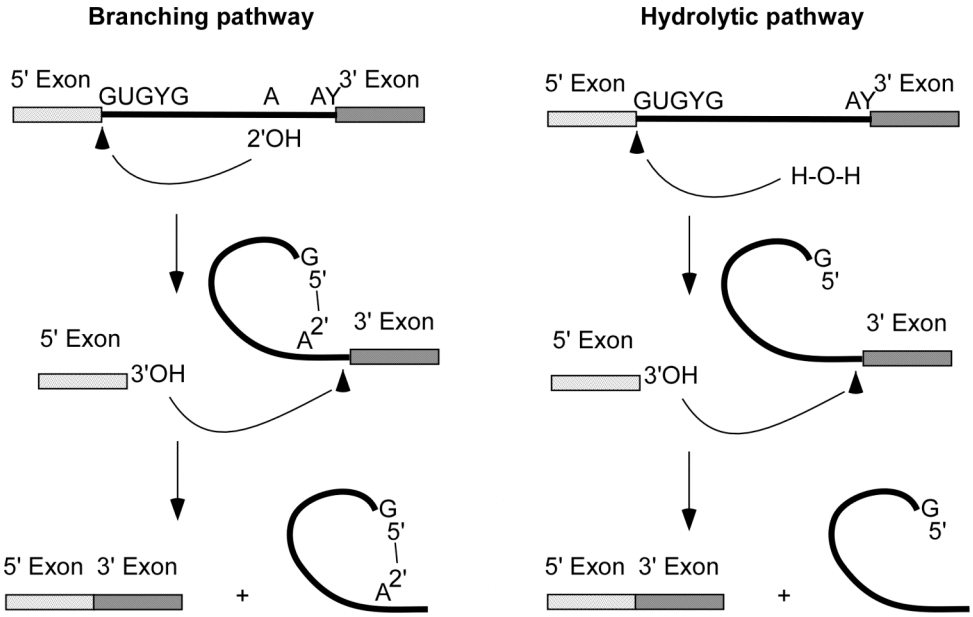



Figure 1.2: Pathways for group II intron splicing.

Group II introns splice via RNA-catalyzed transesterification reactions involving formation of a lariat or linear intermediate. For the branching pathway, the 2' hydroxyl group of the A residue at the branch point in DVI attacks the 5'-splice site to generate a lariat intron attached to the 3'-exon and a free 5'-exon. In the second step, the 3' hydroxyl group of the free 5'-exon attacks the 3'-splice site to generate ligated exons and a lariat intron. In the hydrolytic pathway, a water molecule attacks the 5'-splice site and cleaves the 5'-exon from the intron forming a linear intron with 3'-exon. The second step is the same as the branching pathway and results in a linear intron and ligated exons. The linear and lariat introns are depicted as solid black lines. Exons are shown as boxes. The consensus sequences at the splice sites and the branch point are shown.

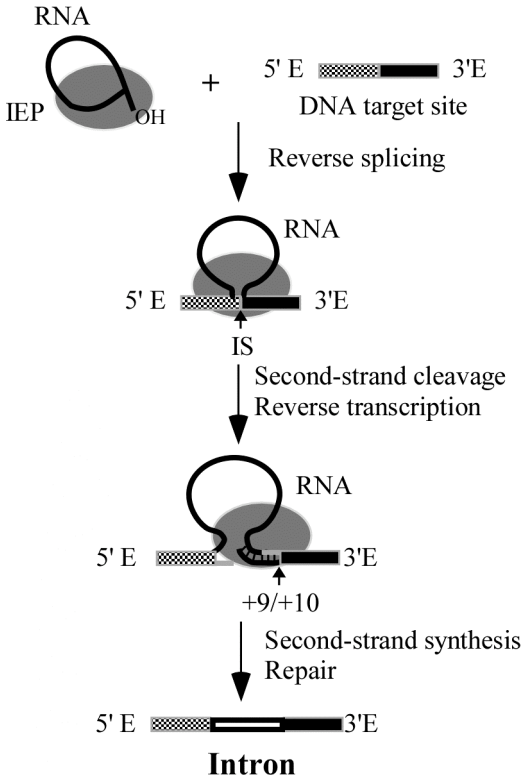



Figure 1.3: Ll.LtrB retrohoming mechanism.

Retrohoming occurs via target DNA-primed reverse transcription (TPRT). The lariat intron RNA in RNPs reverse splices directly into one strand of the DNA target site, while the IEP cleaves the opposite strand between position +9 and +10 of the 3'-exon, and then uses the cleaved 3' end as a primer to reverse transcribe the inserted intron RNA. The resulting intron cDNA is integrated into the recipient DNA via cellular recombination or repair mechanisms. Intron RNA is shown as a black line. DNA target site are shown as gray and black boxes. IEP is the gray oval. 5' E and 3' E denote 5'-exon and 3'-exon in the target site.

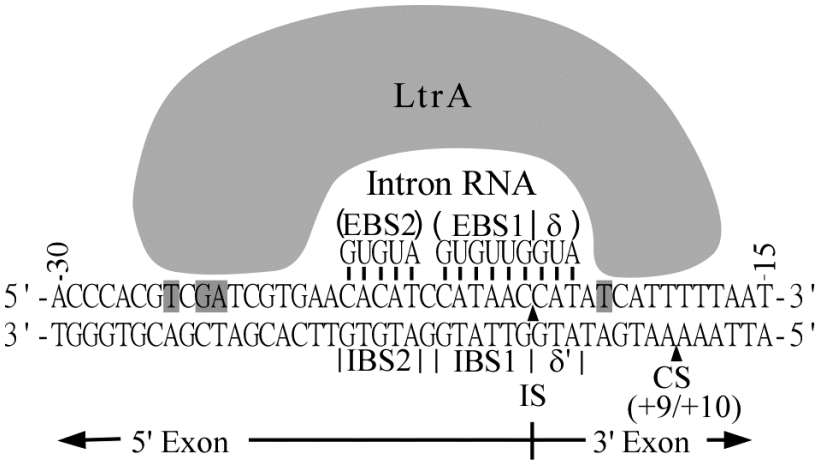



Figure 1.4: DNA target site recognition rules by Ll.LtrB intron.

The DNA target site of Ll.LtrB intron is recognized by both the intron RNA and IEP (LtrA). The LtrA first contacts a small number of bases in the 5' exon region of the DNA target site (T-23, G-21 and A-20, highlighted in gray) via major groove interactions. These base interactions and neighboring phosphate-backbone interactions trigger local DNA melting, enabling the EBS1, 2 and δ sequences in the intron RNA to base pair with the IBS1, 2 and δ' sequences in the DNA target site. The intron RNA directly inserts into one strand of DNA by reverse splicing (IS represents the intron insertion site). The LtrA then cleaves the second strand after recognizing T+5 in the 3'-exon (highlighted in gray). The LtrA cleavage site, denoted CS, is between position +9 and +10 in 3'-exon.

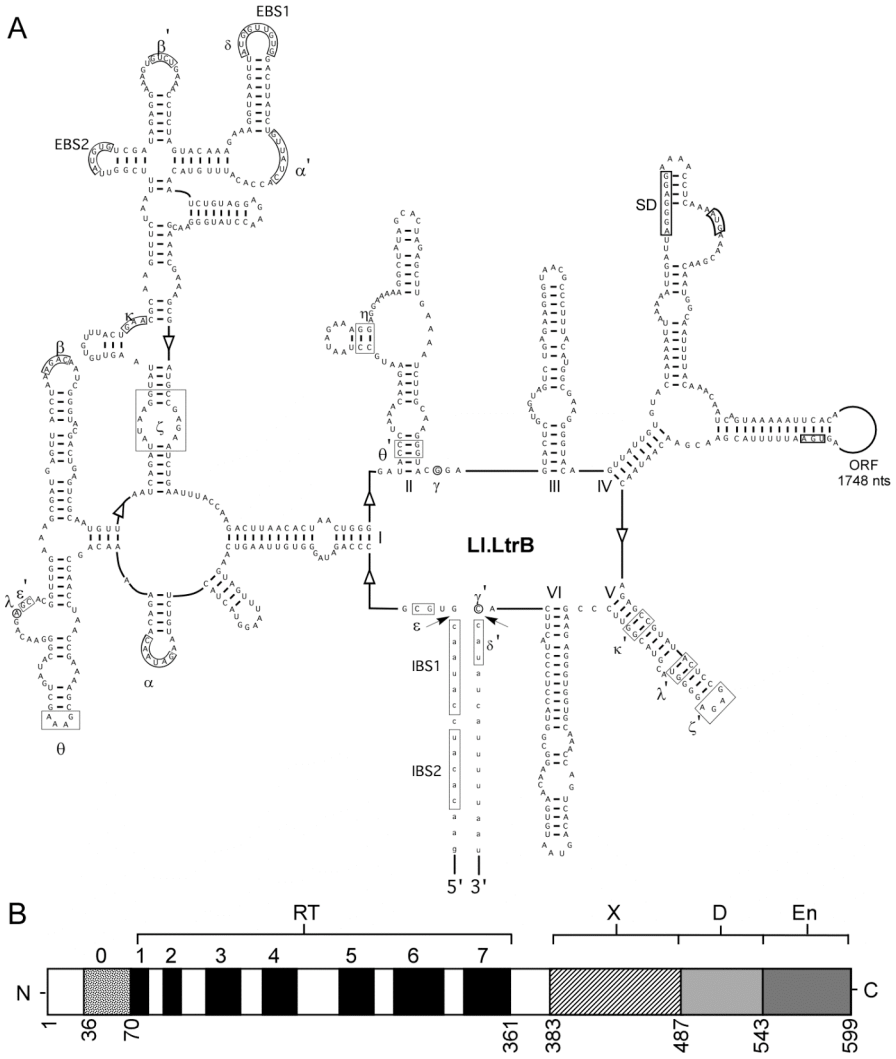



Figure 1.5: Group II intron Ll.LtrB and its IEP (LtrA).

(A) Predicted secondary structure of Ll.LtrB RNA adapted from Michael Karberg's dissertation 2009. The structure consists of six double-helical domains (DI-VI). The 5'- and 3'-splice sites are indicated by arrows. The putative Shine-Dalgarno (SD) sequence, initiation and termination codons of the intron ORF, and sequences involved in tertiary interactions (Greek letters) are boxed. (B) Schematic of LtrA showing conserved domains. RT, containing conserved amino acid sequence blocks 1-7 characteristic of the finger and palm regions of retroviral RTs; X, region associated with maturase activity corresponding in part to the RT thumb; D, DNA binding; and En, DNA endonuclease. 0, region conserved in RTs of non-LTR retroelements. The numbers at the bottom indicate the amino acid position number.

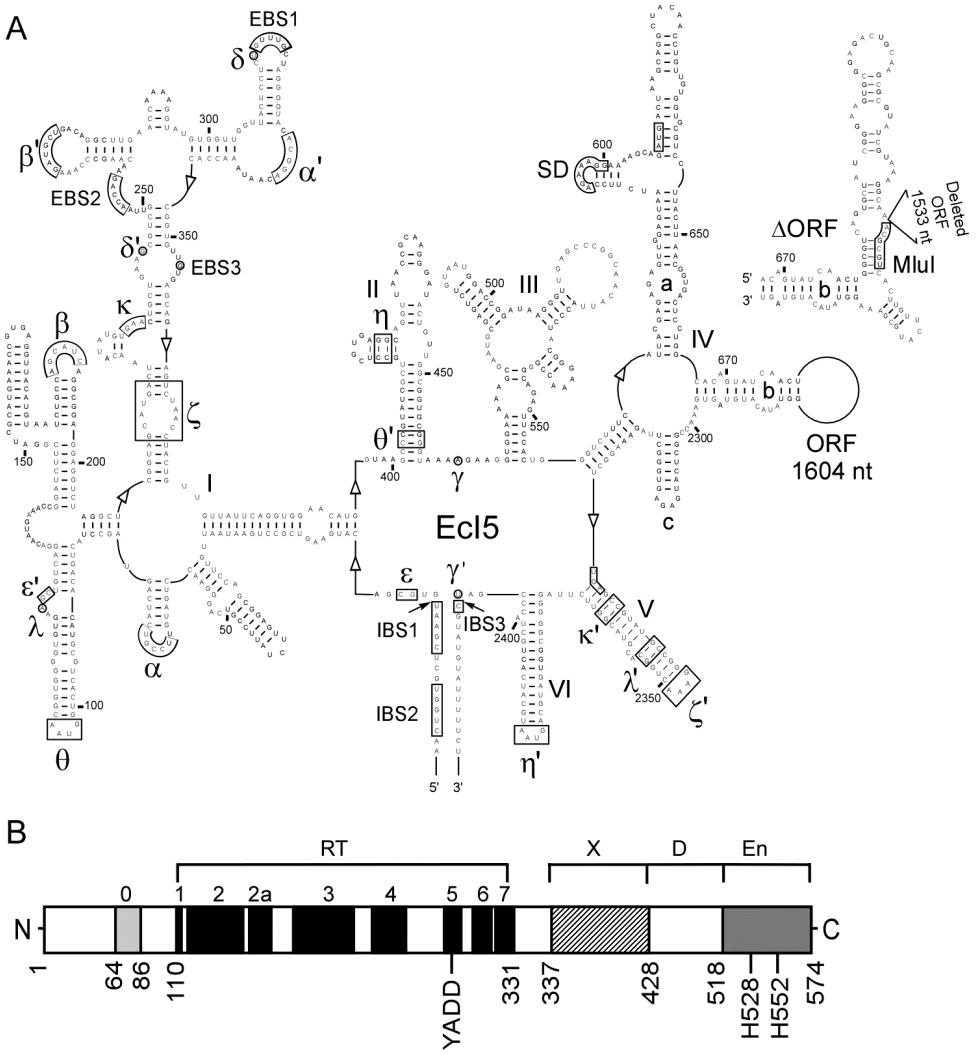



Figure 1.6: Group II intron EcI5 and its IEP.

(A) Predicted secondary structure of EcI5 RNA adapted from Dai and Zimmerly (2002b). The structure consists of six double-helical domains (domain I-VI). The 5'- and 3'-splice sites are indicated by arrows. The putative Shine-Dalgarno (SD) sequence, initiation and termination codons of the intron ORF, and sequences involved in tertiary interactions (Greek letters) are boxed. Numbers indicate nucleotide positions counting from the first nucleotide of the EcI5 intron. (B) Schematic of the EcI5 IEP showing conserved domains. RT, containing conserved amino acid sequence blocks 1-7 characteristic of the finger and palm regions of retroviral RTs; X, region associated with maturase activity corresponding in part to the RT thumb; D, DNA binding; and En, DNA endonuclease. 0, region conserved in RTs of non-LTR retroelements. The numbers at the bottom indicate the amino acid position number.

## Chapter 2: Cellular distribution of LtrA protein in *E. coli*

The *Lactococcus lactis* Ll.LtrB intron has been used as a model system for studying mobile group II introns. Ll.LtrB is highly mobile not only in *L. lactis*, but also in a variety of other Gram-negative and Gram-positive bacteria, including *E. coli*, where its mobility mechanism has been analyzed by using the extensive genetic and biochemical methods available for that organism (Cousineau et al. 1998; Mohr et al. 2000; Karberg et al. 2001; Yao and Lambowitz 2007; Malhotra and Srivastava 2008; Rodriguez et al. 2008). The broad host range of the Ll.LtrB intron reflects that the ribonucleoproteins (RNPs) containing the intron-encoded protein (IEP) and intron RNA can carry out the RNA splicing, reverse splicing, and target DNA-primed reverse transcription reactions without relying on host proteins, while subsequent cDNA integration steps are carried out by common host DNA repair enzymes (Smith et al. 2005; Coros et al. 2008; Zhao et al. 2008).

Recent studies raised the possibility that Ll.LtrB might be associated with the cell membrane at the cell poles. First, experiments using an Ll.LtrB intron library with randomized EBS2, EBS1 and δ sequences showed that although the intron inserted at sites throughout the *E. coli* genome, insertion sites were strongly clustered around the bidirectional replication origin (*oriC*). In a sampling of ~100 insertion events, 57% of the insertion sites were within 5% of the chromosome on either side of *oriC* (Zhong et al. 2003). A separate study of Ll.LtrB retrotransposition in a different strain of *E. coli* showed insertion sites clustered in both the *oriC* and *ter* regions (Ichiyanagi et al. 2002). In *E. coli*, the *oriC* and *ter* regions are localized near the cellular poles during much of the cell cycle (Niki and Hiraga 1998; Zhong et al. 2003). Thus, it is possible that the

clustering of Ll.LtrB insertions sites in the *oriC* and *ter* regions reflects the localization of Ll.LtrB RNPs near the cellular poles.

Zhao and Lambowitz (2005) using LtrA-GFP fusions and immuno-fluorescence microscopy showed that LtrA does in fact localize to cellular poles in both *E. coli* and *L. lactis* and appears to be apposed to the cellular membrane in these regions. The polar localization of LtrA occurred with or without the expression of Ll.LtrB intron RNA, was observed over a wide range of cellular growth rates and expression levels, and was independent of replication origin function (Zhao and Lambowitz 2005). Additionally, the LtrA expression in *E. coli* interfered with the polar localization of co-expressed *Shigella flexneri* IcsA protein, suggesting competition for a common localization determinant (Lambowitz et al. 2005; Zhao and Lambowitz 2005).

Finally, Zhao et al. (2008) using cell microarrays to screen a transposon-inserted library isolated *E. coli* mutants in which LtrA shows a more diffuse localization patterns away from the poles, and Ll.LtrB shows a more uniform genomic distribution of insertion sites in such mutants. These findings indicate that polar localization of LtrA accounts for the preferential insertion of the Ll.LtrB intron in the origin and terminus regions of the *E. coli* chromosome. Interestingly, all of the *E. coli* mutants in which LtrA shows more diffuse localization were found to accumulate intracellular polyphosphate, which appears to bind LtrA and delocalize it from the poles (Zhao et al. 2008).

Together, the above findings raise the possibility that LtrA might be localized by binding to a receptor molecule in the membrane at the cell poles. At present, little is known about receptors that might mediate polar localization of proteins in *E. coli*.

Based on these LtrA localization data, I began a screening for host factors interacting with LtrA in *E. coli*. To identify the host factors associating with LtrA, cells

were treated with cross-linking reagent formaldehyde and disrupted by sonication. The disrupted cells were subjected to centrifugation at 10,000×g for 15 min. The supernatant contains soluble proteins from cell lysate, and the pellet contains membranes, insoluble proteins and unbroken cells. In the non-crosslinked cells, the majority of LtrA can be found in the supernatant. However, after cross-linking, LtrA was no longer found in the supernatant. It could only be detected in the pellet. This finding suggested that LtrA might be cross-linked to large aggregates or associated with membranes or other large structures.

To further understand this phenomena and the mechanism of LtrA polar localization, I investigated the subcellular localization of LtrA protein in *E. coli*. By taking advantage of the well-established cell fractionation procedures in *E. coli*, I characterized the distribution of LtrA protein in different cellular fractions (cytoplasm, inner membrane, outer membrane and periplasm). A high proportion of LtrA protein (36%) was found in the cellular membrane fraction, and the majority of LtrA protein co-purified with the outer membrane after ultracentrifugation in a sucrose gradient. *In vitro* RNA splicing assays demonstrated that the LtrA protein within the membrane fractions was active in promoting the splicing of Ll.LtrB intron precursor RNA.

### 2.1 CONSTRUCTION OF PLAC-LTRA

To study the intracellular localization of LtrA in *E. coli*, I cloned the LtrA ORF downstream of a *lac* promoter in the place of GFP in the vector pGFPuv (Clontech). The resulting plasmid was named plac-LtrA. Previously, it was shown that the polar localization of LtrA was unaffected by co-expression of the intron RNA (Zhao and Lambowitz 2005). Thus I analyzed the subcellular distribution of LtrA expressed in *E. coli* from plac-LtrA without co-expressing Ll.LtrB RNA. Because over-expression of

foreign proteins can result in the formation of inclusion bodies (Strandberg and Enfors 1991), I maintained a low level of expression of LtrA by basal transcription from the *lac* promoter. To do this, the *E. coli* HMS174(DE3) containing plac-LtrA was grown in LB with ampicillin at 37ºC without isopropyl β–D-1-thiogalactopyranoside (IPTG) induction.

### 2.2 LTRA IS PRESENT IN SPHEROPLASTS

*E. coli* periplasm and spheroplast were separated by a lysozyme-EDTA-cold osmotic shock procedure (Osborn et al. 1972a; Osborn and Munson 1974). Freshly grown bacterial cells were treated with sucrose, lysozyme and EDTA on ice to disrupt the cell wall and peptidoglycan. After centrifugation, the supernatants contained the periplasmic fraction, and the spheroplasts formed a pellet. Using this method, conversion to spheroplasts was generally greater than 98%, as judged by phase contrast microscopy (Osborn et al. 1972b).

To check the distribution of LtrA, proteins from the periplasmic fraction, spheroplasts and whole cells were separated in a 0.1% sodium dodecyl sulfate (SDS)/ 7.5% polyacrylamide gel and analyzed by Western blotting using antibodies against LtrA and cellular marker proteins. The amount of each fraction loaded in gel was normalized to the same percentage of volume from each fraction, as described in the Methods section 2.4.2. As shown in Figure 2.1A, almost all of the LtrA from the whole cell was found in the spheroplast fraction and virtually none was found in the periplasmic fraction.

To assess the purity and enrichment of each cellular fraction in this and other experiments, I checked for the presence of different marker proteins by Western blotting. TatA, β-galactosidase, and DsbA proteins were used as markers for cellular membranes, cytoplasm and periplasm, respectively. TatA is one of three integral membrane proteins

in the twin-arginine transport (Tat) system, which translocates folded proteins across the bacterial cytoplasmic membrane (Lee et al. 2006). Studies in *E. coli* using a TatA-GFP fusion protein showed that TatA is present inner membrane, often with higher abundance at the poles (Berthelmann and Bruser 2004). *E. coli* β-galactosidase (*lacZ*) is a cytoplasmic enzyme and is widely used as a cytoplasmic marker (Gerard et al. 2002). The DsbA protein is a periplasmic enzyme which catalyzes disulfide bond formation (Belin and Boquet 1994). In Figure 2.1, all of the marker proteins showed the expected distribution between periplasm and spheroplasts, except for a light band corresponding to the membrane marker TatA in the periplasmic fraction. This light band could be due to a small amount of cross-contamination or leakage from the spheroplast gel lane during loading.

### 2.3 LTRA IS PRESENT IN THE MEMBRANE FRACTION

A spheroplast is a cell from which the cell wall has been almost completely removed. It still contains membrane and cytoplasm. To further examine the LtrA localization within the spheroplast, the cytoplasm and membrane were further separated from the spheroplast. Spheroplasts were lysed by using a cold French press to disrupt the cellular membrane. This method has the advantage of being more reproducible than other cell lysis methods (Nikaido 1994). The disrupted cells were centrifuged twice at 1200×g for 10 min to pellet any remaining intact cells. Then, the cell lysate was ultracentrifuged at 144,748×g for 2 h at 4ºC to yield a cytoplasmic fraction and a membrane pellet. For the cytoplasmic fraction only the top layer of supernatant was collected for analysis in order to avoid contamination by the membrane fraction. After removing of the rest of the supernatant, the pellet was rinsed twice with 50 mM Tris-HCl pH 8.0 to remove any

residual cytoplasm, and the membrane pellet was resuspended in 50 mM Tris-HCl pH 8.0.

To check the LtrA distribution, normalized amounts of spheroplast, cytoplasmic, and membrane fractions were analyzed by 0.1% SDS/ 7.5% polyacrylamide gel electrophoresis and Western blotting (see Methods for normalization). As shown in Figure 2.1B, more than 80% of LtrA was present in the membrane fraction with only a small amount of LtrA detected in the cytoplasm. Western blotting to detect the TatA and β-galactosidase marker proteins confirmed them to be present mainly in the membrane and cytoplasmic fractions, respectively, although a small amount of TatA was detected in the cytoplasmic fraction.

### 2.4 DISTRIBUTION OF LTRA PROTEIN IN THE INNER AND OUTER MEMBRANES

The association of LtrA protein with membrane or a membrane receptor might explain its bipolar localization in *E. coli*. For gram-negative bacteria, the cell membranes consist of three morphologically distinguishable layers: an inner membrane surrounding the cytoplasm; a murein (peptidoglycan) layer external to the cytoplasmic membrane; and an outer membrane at the external surface of the cell (Murray et al. 1965). As a lipid bilayer membrane, the outer membrane behaves similarly to the inner membrane during cell fractionation. However, separation of the two membranes is possible because of differences in composition. One method is based upon their different buoyant densities. Another method is based on their different lipid compositions. Most of the lipids in the outer membrane are lipopolysaccharides (LPS), which have very different properties from those of the glycerophospholipids comprising the bilayer of the inner membrane. One can selectively solublize the inner membrane without dissolving much of the outer membrane. Finally the presence of LPS makes the outer membrane an intrinsically

unstable structure. By destabilizing the membrane, the outer membrane are shed off as blebs or vesicles, which can be collected (Nikaido 1994).

Here I separated the inner and outer membrane on the basis of their different buoyant densities by equilibrium density gradient centrifugation. After breaking the cells, the membranes were separated by centrifugation through a sucrose density gradient (Osborn and Munson 1974; Nikaido 1994). Four discrete membrane bands appeared from the top to bottom of the gradient (Figure 2.2 and Methods). Bands L1 and L2 have been characterized as inner membrane, H as outer membrane, and the minor band, M, at an intermediate density as envelope fragments. The presence of band M is influenced by several factors such as growth medium and ionic strength (Osborn et al. 1972a). The separation of inner membrane fragments into two bands appears to reflect heterogeneity in the degree of contamination by outer membranes. To evaluate the efficiency of separation, I assessed the separation of inner and outer membrane by checking for the presence of the inner membrane marker protein, TatA, using Western blotting (Figure 2.2), and other proteins detected by Coomassie staining of SDS-polyacrylamide gels (Figure 2.3).

Figure 2.2 shows that the TatA protein was present mainly in fraction 29 to 39 corresponding to the two inner membrane bands L1 and L2. A Coomassie stained 2% SDS/ 7.5% polyacrylamide gel (Figure 2.3) shows that two major bands of molecular weight ~32 kDa and a series of more rapidly migrating bands are specific to the outer membrane. As expected, the fractions with outer membrane proteins in Figure 2.3 overlapped the H band in Figure 2.2 (fractions 45 to 62). The SDS-PAGE profiles of L1 and L2 were indistinguishable from each other and contained a number of high molecular weight proteins (>50 kDa) in a Coomassie stained SDS-polyacrylamide gel (Figure 2.3).

After confirming the separation of inner and outer membranes, I checked the distribution of LtrA protein by Western blotting using antibody against LtrA. Surprisingly the LtrA was mainly found in outer membrane fractions (band M and H; fractions 40 to 58 in Figure 2.2), but with some trailing into the inner membrane fraction L2 (fractions 29 to 39 in Figure 2.2).

Figure 2.4 summarizes the distribution of LtrA protein in the membrane fractions, the total protein concentration of each fraction, and the distribution of TatA and outer membrane proteins in the membrane fractions. The relative amount of LtrA and TatA proteins in each fraction was calculated from the band intensity of LtrA and TatA in the Western blots of Figure 2.2, as described in Methods. The relative amount of outer membrane proteins in each fraction was calculated from the band intensity of the two major outer membrane protein bands in Figure 2.3. The total protein concentration in each fraction of Figure 2.4 reflects the actual protein concentration. The amount of LtrA, TatA and outer membrane proteins in Figure 2.4 was normalized to a scale 0 to 4 to fit in the same plot. As shown in Figure 2.4, the peak of LtrA protein (blue line) overlaps that of outer membrane proteins (green line) and the H band ( red line).

### 2.5 RELATIVE DISTRIBUTION PROFILE OF LTRA AMONG CELLULAR FRACTIONS

The cell fractionation experiments above show LtrA to be present in different cellular fractions, including cytoplasm, inner and outer membranes. I quantified the amount of LtrA in the different cellular fractions (Figure 2.1C). The amount of each fraction loaded on the gel was normalized, as described in Methods, and the purity of each fraction was confirmed by the presence of different marker proteins (data not shown). Based on the analysis of band intensity by Bio-Rad Quantitative One software (Bio-Rad), 83% of LtrA protein was present in the total membrane fraction compared to

17% in the cytoplasm. 30% of LtrA protein was found in the outer membrane fractions and only 6% of LtrA protein was found in the inner membrane. A substantial amount of LtrA protein (47%) was found in the pellet after sucrose gradient centrifugation, possibly reflecting the lower solubility of LtrA in cell lysis buffer or aggregation after dissociation from cellular membranes.

Because LtrA has a calculated pI of 9.60, it is positively charged at physiological pH and has a high non-specific binding affinity for nucleic acid (Singh and Lambowitz 2001; Aizawa et al. 2003). Therefore, I next addressed the question of whether the distribution profile of the LtrA protein in different fractions is altered by treatment with nucleases. In Figure 2.5A, I repeated the cell fractionation using the same procedures as in Figure 2.1C, except for the addition of protease inhibitor, RNase-free DNase I, RNase A or different combinations prior to lysis of the spheroplasts (see Methods). As shown in Figure 2.5A, with addition of protease inhibitors only, LtrA was detected in both the cytoplasmic and membrane fractions. With the addition of DNase I, however, the amount of LtrA in the cytoplasmic fraction decreased and with addition of RNase A or RNase A and DNase I together, the amount of LtrA in the cytoplasmic fraction decreased to undetectable levels. Notably, in the presence of nucleases the amount of LtrA in the pellet increased. Meanwhile, the amount of LtrA in the other fractions was not significantly affected by the above treatments. These results suggest that soluble LtrA might be stabilized in the cytoplasm by binding to nucleic acids, especially RNA, and becomes insoluble after the nucleic acids are digested.

Figure 2.5B shows a control experiment in which purified LtrA protein was added directly to HMS174(DE3) spheroplasts. After lysis, no LtrA was detected in the cytoplasmic fraction, or in the inner or outer membrane fractions. Instead, almost all of

the added LtrA became insoluble and was found in the pellet after centrifugation. This experiment provides evidence that the endogenous LtrA found in cytoplasmic and membrane fractions is physiologically relevant and does not result from non-specific interactions of soluble protein with membrane fractions after cell breakage.

Together, the above experiments show that endogenous LtrA is present in both the cytoplasmic and membrane fractions, that endogenous LtrA found in the cytoplasmic fraction may contain bound RNA and DNA that keeps the protein soluble after cell breakage, and that the association of endogenous LtrA with cellular membranes was unaffected by nuclease treatment.

### 2.6 LtrA in the membrane can promote Ll.LtrB intron splicing *in vitro*

I next determined whether the LtrA present in the membrane fraction is functional by assaying whether it could stimulate Ll.LtrB precursor RNA splicing *in vitro*. A splicing assay was done by using an *in vitro* transcript precursor RNA containing the Ll.LtrB-ΔORF intron and short flanking exons, as described previously (Saldanha et al. 1999). LtrA from membrane fractions was prepared by sucrose gradient centrifugation, with Western blotting for LtrA and marker proteins used to assess the enrichment and purity of the fractions. In the splicing experiment, only fractions representing the peaks of inner and outer membranes were used (fractions 31 to 34 and 54 to 55, respectively). The amount of LtrA in these fractions was quantified by Western blotting using purified LtrA protein as a standard. In the splicing assay, 100 nM of Ll.LtrB precursor RNA and 20 nM of LtrA protein were used in each reaction. The RNA was renatured by heating to 55ºC and slowly cooling to 37ºC, and the splicing reaction was done at 30ºC for 30 min.

In the experiment of Figure 2.6, the splicing activity of purified LtrA protein and LtrA from inner and outer membrane fractions were checked at different NaCl

concentration (0, 50, 100 and 500 mM). The purified LtrA protein stimulated the splicing of Ll.LtrB precursor RNA at NaCl concentration from 0 to 500 mM, with the highest activity at 500 mM NaCl, as found previously (Saldanha et al. 1999). The LtrA protein in the inner membrane fraction also stimulated RNA splicing at all salt concentrations with the highest activity at 500 mM NaCl. However the splicing activity of LtrA in the inner membrane fraction was lower than that of purified LtrA at 500 mM NaCl. The splicing activity of LtrA in the outer membrane fraction was lower than that of either purified LtrA or LtrA in the inner membrane fractions and was detected only at 500 mM NaCl. In control experiments, the boiled inner and outer membrane fractions did not stimulate Ll.LtrB RNA splicing, suggesting that the observed splicing activity is due to LtrA protein and not to heat-resistant factors, such salts or lipids in the membrane fractions.

The above data demonstrated that LtrA bound to membrane can promote the Ll.LtrB precursor RNA splicing. I next investigated whether there were other splicing factors in the membrane that increase the splicing activity of LtrA in the absence of salt or at a salt concentration closer to physiological conditions (100 mM NaCl). For this purpose, I carried out the same splicing assay with purified LtrA protein in the combination with the inner and outer membrane fractions from wild-type HMS174(DE3) cells not expressing LtrA protein.

As shown in Figure 2.7, at 0 mM NaCl, the splicing activity of purified LtrA protein was barely detectable (lane 2). The inner membrane fraction by itself did not promote the splicing of the Ll.LtrB precursor RNA (lane 3). However, combining 20 nM of purified LtrA with the inner membrane fraction resulted in a higher level of splicing than with purified LtrA protein alone (lane 4, compare with lane 2). The inner membrane fraction contains nucleases that degrade precursor RNA, so it is possible that the activity

of LtrA plus inner membrane fraction is higher than appears from the gel. Notably, the addition of purified LtrA protein to boiled inner membrane showed somewhat higher RNA splicing activity than that of purified LtrA or LtrA with inner membrane (lane 6, compared to lanes 2 and 4). This finding suggests that heat-resistant factors present in the inner membrane fraction may stimulate LtrA splicing activity under low salt conditions. Similar results were obtained for the outer membrane fraction. The outer membrane fraction did not by itself stimulate Ll.LtrB RNA splicing (lane 7). However, LtrA together with outer membrane fraction or boiled outer membrane fraction stimulated splicing of the Ll.LtrB RNA to somewhat higher levels than were seen with purified LtrA protein alone (lane 8 and 10, compare to lane 2).

In splicing assays with 100 mM NaCl, the splicing activity of purified LtrA was higher than that without NaCl (lane 2 and 11). Surprisingly, under these conditions, there was detectable splicing activity with the inner and outer membrane without LtrA protein (lane 12 and 16). This activity was abolished after boiling the inner and outer membrane fractions (lane 14 and 18). These findings suggest that heat sensitive factors in the inner and outer membranes can by themselves support Ll.LtrB RNA precursor splicing at 100 mM NaCl, even without LtrA protein. Again, the splicing activity of Ll.LtrB with those unknown membrane factors might be higher than what was observed because a significant amount of RNA precursor was degraded in the reactions containing membrane fractions. RNA degradation may also contribute to the lower splicing activity of purified LtrA at 100 mM NaCl in the presence of inner membrane (lane 13) or outer membrane (lane 17) compared to that of LtrA only (lane 11).

In summary, the splicing experiments using purified LtrA in combination with inner or outer membrane fractions gave two potentially interesting findings. First, the

experiments raise the possibility that the membrane fractions contain unknown heat-sensitive splicing factors that can support Ll.LtrB intron splicing without LtrA at a condition close to physiological salt concentrations (100 mM NaCl). Second, in the absence of NaCl, boiled inner or outer membrane fractions may increase the splicing activity of purified LtrA. These findings raise the possibility that association of LtrA with *E. coli* membranes enhances its biological activity. Further experiments in which nuclease activities are inhibited will be required to confirm the increase in splicing activity for membrane fractions.

## 2.7 DISCUSSION

The results presented in this chapter showed that after cell breakage and fractionation, LtrA protein is present in *E. coli* cytoplasmic, inner and outer membrane fractions. Surprisingly, a high proportion of LtrA is found in the outer membrane fractions after sucrose density gradient centrifugation. These experiments, however, do not distinguish whether LtrA is bound to the outer membrane or is merely co-purifying with it. One possibility is that LtrA is bound to a specific dense fraction of inner membranes that co-purifies with the outer membrane. The LtrA protein recovered in the cytoplasmic fraction appears to be associated with nucleic acids, mainly RNA, whose removal causes the protein to become insoluble. Although splicing assays showed that the LtrA protein associated with inner and outer membrane was active in stimulating the splicing of Ll.LtrB precursor RNA, the biological function of LtrA in different cellular fractions remains to be analyzed in detail. Nevertheless, the binding of LtrA to a pole localized membrane receptor could account for the polar localization of LtrA in *E. coli*.

An unexpected finding was that *E. coli* membrane fractions may contain heat-sensitive factors that can by themselves support Ll.LtrB precursor RNA splicing in the

absence of LtrA at lower salt concentrations (0 and 100 mM NaCl). The heat sensitivity of these factors is consistent with the possibility that they are proteins. Additionally, purified LtrA protein appears to splice more efficiently at low salt concentrations in the presence of inner membrane or boiled inner membrane, suggesting that membrane components, such as lipids, may enhance its activity. The stimulatory effects of membrane fractions in these experiments may be underestimated due to the presence of RNases that degrade precursor RNA, and future experiments with better control of RNA degradation (e.g., by addition of more RNase inhibitor) would increase the accuracy of the assays. It will also be of interest to assay the effect of membranes on other LtrA activities, such as reverse transcriptase, DNA endonuclease, and reverse splicing.

## 2.8 METHODS

### 2.8.1 *E. coli* strain and recombinant plasmids

The *E. coli* strain HMS174(DE3) strain was used to express LtrA protein. plac-LtrA is derived from pGFPuv (Clontech). To construct plac-LtrA, pGFPuv was first digested with HindIII and SpeI to remove the GFP gene. LtrA was amplified from pIMP-lp (Saldanha et al. 1999) using primers LtrA184s (5'-GCGGCGaagcttGATGAAACCAACAATGG-CAATTTTAG-3') and LtrA185a (5'-GCGGCGactagtTCACTTGTGTTTATGAAT-CACGTG-3'). LtrA184s appends a HindIII site, and LtrA185a appends a SpeI (restriction enzyme sites are in lower case letters in the primer sequences). The PCR product was digested with HindIII and SpeI and swapped for the GFP gene in the pGFPuv vector. plac-LtrA contains a β-lactamase gene, which confers ampicillin resistance. HMS174(DE3) transformed with plac-LtrA was grown in LB medium with 100 μg/ml ampicillin.

### 2.8.2 Bacterial cell fractionations

*E. coli* HMS174(DE3) cells were electro-transformed with plasmid plac-LtrA and grown in the presence of ampicillin overnight at 37ºC. The overnight culture was then subcultured 1:100 into fresh LB media with ampicillin and grown to an optical density at 600 nm ($O.D._{600}$) of 0.5. Cellular fractions were prepared by a lysozyme-EDTA-cold osmotic shock procedure (Osborn et al. 1972a).

***Preparation of periplasmic fraction***

4 ml of cells ($O.D._{600}$= 0.5) was harvested at 4ºC by centrifugation at 6000×g for 7 min. The cell pellet was then resuspended in 2 ml freshly made fractionation buffer (18% sucrose, 50 mM Tris-HCl pH 8.0, 1 mM $CaCl_2$), and the cells were centrifuged again at 6000×g for 7 min at 4ºC. 1 ml fractionation buffer was added to resuspend the cell pellet, and a 250 $\mu$l of aliquot was saved as a whole-cell fraction. 0.5 ml of the whole-cell fraction was added to a tube containing 1 µl 0.5 M EDTA and 5 µl of 100 µg/ml lysozyme and incubated on ice for 30 min to form spheroplasts. Following centrifugation at 6000×g for 5 min at 4ºC, 250 µl of supernatant was saved as the periplasmic fraction. The pellet was then resuspended in 500 µl fractionation buffer and saved as the spheroplast fraction.

***Preparation of cytoplasmic and membrane fractions***

50 ml of culture ($O.D._{600}$= 0.5) was harvested by centrifugation at 6000×g for 7 min. The cells were then resuspended in 6 ml of 50 mM Tris-HCl, pH 8.0 and centrifuged again to remove any contaminants from the media. After the centrifugation, the cells were resuspended in 6 ml of 50 mM Tris-HCl, pH 8.0, 0.5 M sucrose, 2.5 mM EDTA and 0.6 mg/ml lysozyme. The samples were left on ice for 30 min to disrupt the cell wall and peptidoglycan. Spheroplasts were then pelleted at 17,000×g for 10 min at 4ºC, and the

supernatants were collected as the periplasmic fraction. The spheroplasts were resuspended in 3 ml of 50 mM Tris-HCl, pH 8.0, with a few flakes of DNase I (Roche) and one Complete, Mini, EDTA-free Protease Inhibitor Cocktail Tablet (Roche). The cells were then disrupted by passing twice through a cold French press (Carver) at a pressure of 20,000 lb/in$^2$. Intact bacteria were removed by centrifuging twice at 1200×g for 10 min at 4ºC. The supernatant, containing soluble proteins (cytosolic), insoluble proteins and membranes, was subjected to ultracentrifugation at 144,748×g for 1.5 h at 4ºC to pellet membranes and insoluble proteins. The supernatant was collected as the cytoplasmic fraction. Membrane and insoluble proteins were washed and resuspended in 0.5 ml of 50 mM Tris-HCl pH 8.0 containing Complete, Mini, EDTA-free Protease Inhibitor Cocktail Tablets (Roche). Enrichment and purity of subcellular fractions were checked by Western blotting for known protein markers.

The amount of each fraction loaded in the gel was normalized to the same percentage of total volume in each fraction. Fractions with large volume were concentrated by speed vacuum and resuspended in SDS-PAGE loading buffer.

***Separation of inner and outer membrane***

The separation of inner and outer membranes is based on published methods (Osborn and Munson 1974; Nikaido 1994). Cells containing plac-LtrA were grown overnight at 37ºC. The overnight culture was inoculated into 1 liter of fresh LB medium containing ampicillin and grown to O.D.$_{600}$= 0.6-0.8. After this step, cells and buffers were kept cold. The cells were harvested at 6000×g for 7 min at 4ºC and resuspended into 100 ml of buffer K (50 mM triethanolamine, 250 mM sucrose, 1 mM EDTA, pH adjusted to 7.5 by acetic acid). The cells were again pelleted at 6000×g for 7 min at 4ºC and resuspended in 15 ml buffer K with or without 1 mg DNase I, 1 mg RNase A and a

Complete, Mini, EDTA-free Protease Inhibitor Cocktail Tablet (Roche) at 4ºC as indicated for individual experiments. Cell membranes were disrupted by passing twice through a cold French press (Carver) at a pressure of 20,000 lb/in$^2$, and cell envelopes and unbroken bacteria were removed by centrifuging twice at 1200×g for 10 min. The supernatant was subjected to ultracentrifugation at 45,000 rpm in a Beckman 70Ti rotor for 90 min at 4ºC. The membrane pellet was rinsed with fresh buffer K twice to remove any cytoplasmic contamination and resuspended in 3 ml of buffer K with one tablet of Complete, Mini, EDTA-free Protease Inhibitor Cocktail at 4ºC by repeated drawing in and squeezing out through a No.23 syringe needle (Becton Dickinson) (Nikaido 1994). The sucrose step gradients were composed of six layers containing 30%, 35%, 40%, 45%, 50% and 55% sucrose. The step gradients were prepared by layering 5 ml of each sucrose solution starting with the highest concentration at the bottom of the centrifuge tubes. The interfaces between the layers were marked. The membrane suspension was layered over the sucrose step gradient and centrifuged at 25,000 rpm in a SW28 rotor (Beckman). The centrifuge was set at slow acceleration and was stopped without brake. Apparent equilibrium was reached after 30-36 h, but adequate separations were achieved by 18 h.

The gradients were collected by puncturing the bottom of the tube with a piecing device and 60% sucrose was slowly and continuously pumped into the centrifugation tube through the hole to push the solution into a fraction collector. Each fraction contained 600 μl and the fractions were collected from the top to bottom of the centrifugation tube. The protein concentration of each fraction was calculated based on the equation: Protein Concentration (mg/ml) = $1.5 \times O.D._{280} - 0.75 \times O.D._{260}$ (Bollag et al. 1996). After ultracentrifugation, the step gradient was converted into a continuous

gradient. Here the concentration of sucrose was considered to increase continuously. Usually, the inner membrane was found at sucrose concentrations of 24-42%, and outer membrane was found at sucrose concentration of 42-55%.

For preparation of total inner and outer membranes, the separated fractions were directly removed from the centrifuge tube using needles attached to a syringe. Four bands, L1, L2, M and H, usually appeared after centrifugation. The lower, white band (band H) with strong turbidity corresponds to outer membranes with the still associated peptidoglycan layer, while inner membranes appear as a translucent, brownish upper bands (bands L1 and L2). The cytoplasmic membranes are found at the interface between the 24 and 42% sucrose, and the outer membrane at the interface between 42% and 55% sucrose.

To pellet inner membranes, 18 ml of buffer M (100 mM triethanolamine, 2 mM EDTA, pH adjusted to 7.5 by acetic acid) was added to 3.5 ml of collected inner membrane fraction and subjected to ultracentrifugation at 47,600 rpm in a 70.1Ti rotor (Beckman) for at least 18 h at 4ºC. After centrifugation, the brown pellet was resuspended in 300 µl buffer M with protease inhibitors. To pellet outer membranes, ~ 4 ml of outer membrane fraction was diluted into 60 ml of buffer M, and 3 ml of 10% sarkosyl was added. The solution was mixed at room temperature by a constant rate mixer for 20 min and then centrifuged at 10,000×g for 1 h in an SW28 rotor. After centrifugation, the pellet was rinsed with buffer M and resuspended in 150 µl buffer M with protease inhibitor.

### 2.8.3 Western blotting

Cell extracts were subjected to sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) and transferred to PVDF membrane (Bio-Rad) using an

electro-transfer device (Pharmacia Biotech, model#EPS3500). After the transfer, membranes were blocked by 1% dry nonfat milk dissolved in 1×TBS (100 mM Tris-HCl, pH 8.0, 1.5 M NaCl) at room temperature for 1 h. The rabbit anti-thioredoxin and anti-TatA antibodies were obtained from Dr. George Georgiou, UT Austin. The anti-LtrA antibody was the laboratory stock (Zhao and Lambowitz 2005). The anti beta-galactosidase antibody was purchased from Molecular Probes (1:10,000). The secondary antibody used for all Western blotting was alkaline phosphatase conjugated anti-rabbit IgG(Fc) (Promega). BCIP (5-bromo-4-chloro-3-indolyl-phosphate) (Promega) was used in conjunction with NBT (nitro blue tetrazolium) (Promega) for the colorimetric detection of alkaline phosphatase activity.

### 2.8.4 *In vitro* LtrA-promoted Ll.LtrB splicing assay

LtrA-assisted splicing reactions were as described (Matsuura et al. 1997; Saldanha et al. 1999). Substrates for *in vitro* splicing were $^{32}$P-labeled *in vitro* transcripts synthesized with phage T3 RNA polymerase (Stratagene) from pGM–ΔORF linearized with BamHI. *In vitro* transcription was in 50 µl of reaction medium (Caprara et al. 1996) containing 1 µg of template DNA, 0.5 mM each of ATP, GTP, CTP, 0.4 mM UTP, and 5 µCi [ $\alpha$-$^{32}$P]-UTP (3,000 Ci/mmole) for 4 h at 37°C. The DNA template was removed by DNase I treatment (Ambion; 10 units for 15 min at 37°C), and the precursor RNA was purified by electrophoresis on a denaturing 4% polyacrylamide gel. For splicing reactions, the $^{32}$P-labeled precursor RNA (100 nM; ~100,000 cpm) was dissolved in 20 µl of low salt reaction medium containing 5 mM $MgCl_2$, 40 mM Tris-HCl at pH 7.5 and 2 µl of SUPERase•In RNase inhibitor (Ambion, Cat#2694). The amount of NaCl in the reaction buffer was adjusted according to the experiments. The RNA was renatured by incubating to 55°C in a waterbath and then slowly cooling to 37°C. Splicing reactions

were initiated by adding of the 20 nM of purified LtrA protein or membrane fractions containing the equivalent amount of LtrA and incubated at 30°C for times specified for individual experiments. The reactions were terminated by phenol-chloroform-isoamyl alcohol (25:24:1) extraction and ethanol precipitation, and the products were analyzed in a denaturing 4% polyacrylamide gel, which was dried and scanned with a phosphorimager.

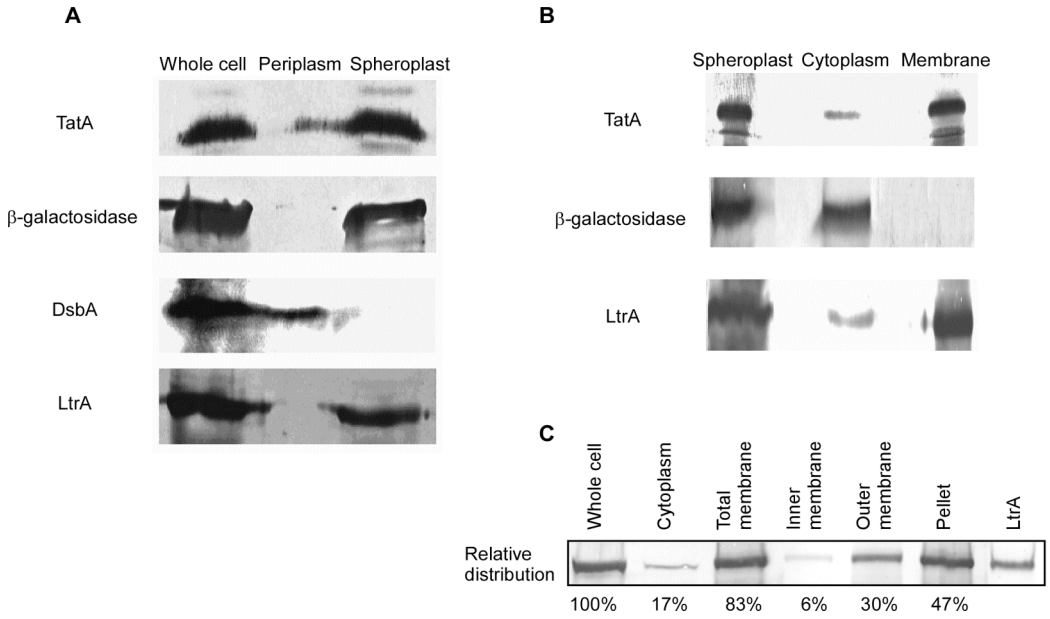



Figure 2.1: The distribution LtrA protein in *E. coli* cellular fractions.

The figure shows Western blot analysis of cellular fractions from *E. coli* HMS174(DE3) expressing LtrA at a low level from the uninduced expression vector, plac-LtrA. The low expression level minimizes aggregation of the LtrA protein. Cellular fractions were prepared, and the amount of each fraction loaded in each lane was normalized as described in Methods. The purity and enrichment of each fraction was confirmed by immunoblotting for different marker proteins. TatA, β-galactosidase and DsbA were used as markers for membranes, cytoplasm, and periplasm respectively. The cellular fractions are indicated on the top of each Western blot, and the antibodies used are shown on the left of the gels. (A) Distribution of LtrA in periplasm and spheroplast fractions. (B) Distribution of LtrA protein in cytoplasmic and membrane fractions. (C) Distribution of LtrA protein among all cellular fractions. The last lane contains purified LtrA protein. The pellet was collected from the sucrose gradient centrifugation step. The relative amount of protein in each lane was calculated from the band intensity in the Western blot using Quantity One software (BioRad).

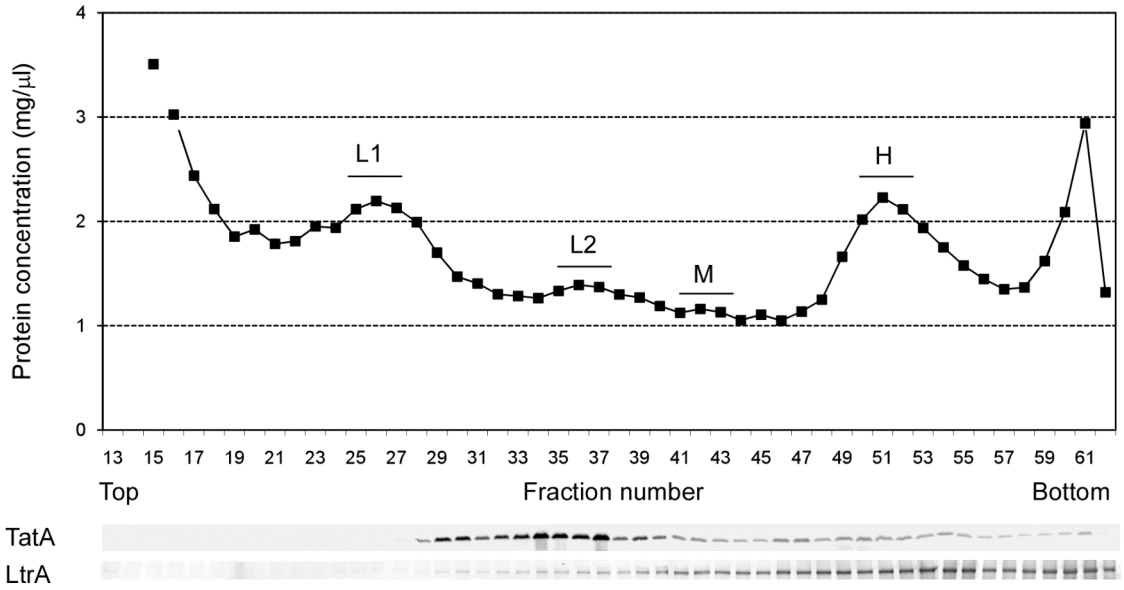



Figure 2.2: Distribution of LtrA protein in *E. coli* HMS174(DE3) inner and outer membranes.

A total membrane fraction was prepared from *E. coli* HMS174(DE3) cells expressing LtrA protein. Inner and outer membranes were then separated by sucrose gradient ultra-centrifugation, as described in Methods. 600 μl fractions were collected from the top to bottom of the centrifuge tube. The absorbance of each fraction at 280 nm and 260 nm was measured by using a Nanodrop (Thermo scientific). The protein concentration was calculated based on the equation: Protein concentration (mg/ml) = $1.5\times O.D._{280} - 0.75\times O.D._{260}$. Fours bands, corresponding from the top to bottom to protein concentration peaks L1, L2, M and H were observed after ultracentrifugation. The x-axis is the fraction number, and the y axis is the total protein concentration in each fraction. The presence of inner membrane protein marker, TatA, and LtrA protein in each fraction was checked by Western blot analysis (shown at bottom). Equal volumes of each fraction were loaded in each lane of the gel used for Western blotting.

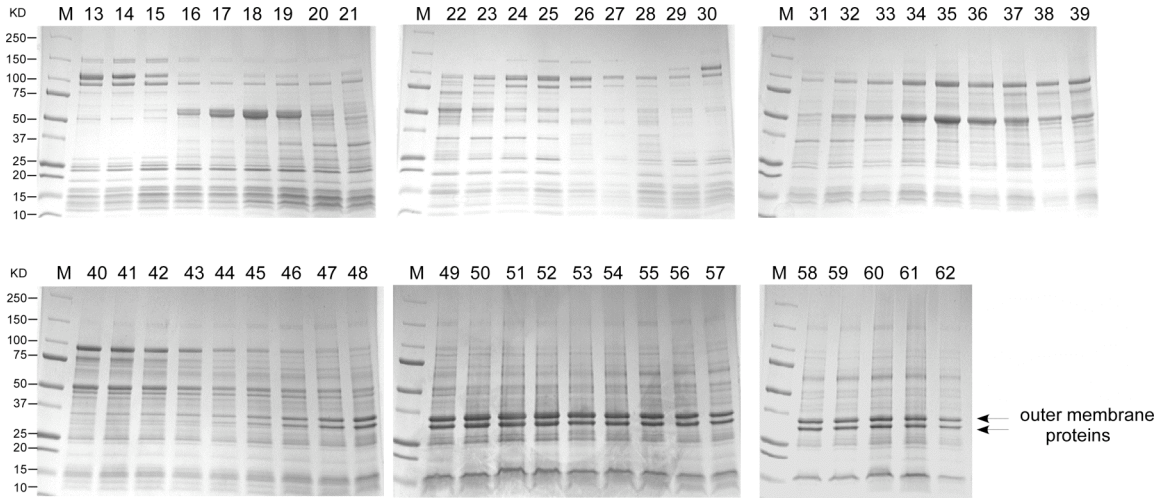



Figure 2.3: The distribution of outer membrane proteins in membrane fractions after sucrose gradient centrifugation.

The protein profile of each fraction in Figure 2.2 was analyzed by electrophoresis in a 2% SDS/ 7.5% polyacrylamide gel. Equal volumes of fraction 13 to 62 collected from sucrose gradient centrifugation were loaded. The protein gels were then stained with Coomassie blue. The outer membrane proteins, indicated with arrows, were present in fractions 45 to 62, with a peak around fraction 51. M, protein ladder (Bio-Rad, Prescision Plus Protein Prestained Standards, All blue). Fraction numbers are indicated above each lane.

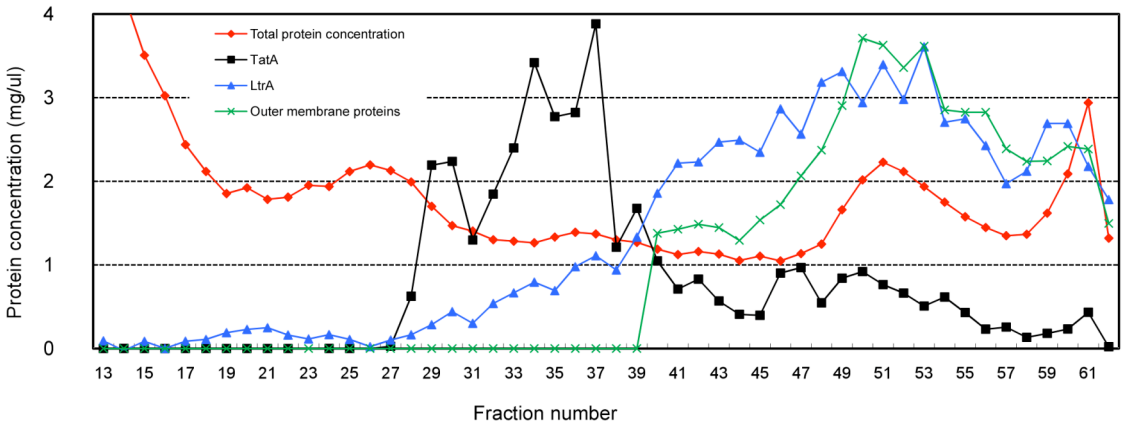



Figure 2.4: Protein concentrations and relative distribution of LtrA, TatA and outer membrane proteins among sucrose gradient fractions.

The protein concentration (red line) of each fraction was calculated as described in Methods, and plotted against fraction number. The relative amount of LtrA protein (blue line) and TatA protein (black line) in each fraction were calculated from the band intensity from the Western blots in Figure 2.2. The amount of outer membrane proteins in each fraction was calculated from the band intensity in Figure 2.3. Band intensities were determined by Quantitative One software (Bio-rad). The x-axis indicates the fraction number. The y-axis shows the amount of protein in each fraction. The relative amounts of LtrA, TatA and outer membrane proteins in the fractions were normalized to fall within the range 0 to 4.

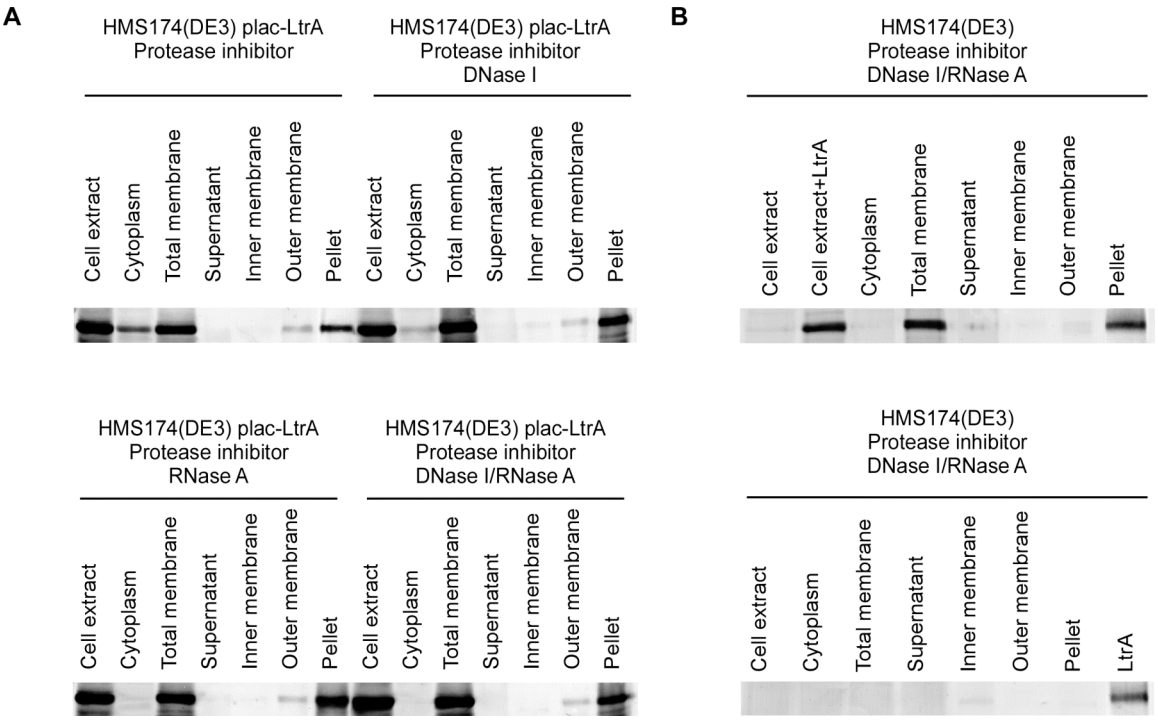



Figure 2.5: Effect of nuclease treatment on the distribution of LtrA protein among *E. coli* cellular fractions.

Cellular fractions were prepared as described in Methods from HMS174(DE3) cells expressing LtrA, or HMS174(DE3) cells with purified LtrA protein added to the extract prior to fractionation. (A) Effect of nuclease treatment on the LtrA protein distribution. Spheroplasts of *E. coli* HMS174(DE3) expressing LtrA were prepared and resuspended in buffer containing protease inhibitor, protease inhibitor with DNase I, RNase A or both, as indicated on the top of each gel. Cellular fractions were then prepared, and the relative amount of LtrA protein in each fraction was analyzed by Western blotting using anti-LtrA antibody. The amount of each fractions loaded on the gel was normalized, as described in Methods. The purity of each cellular fraction was assessed by Western blotting for protein markers TatA, β-galactosidase, and DsbA for cellular membranes, cytoplasm and periplasm, respectively. The distribution of outer membrane proteins was also checked (data not shown). (B) Control showing subcellular fractionation after addition of purified LtrA protein to an *E. coli* HMS174(DE3) cell extract. The purified LtrA protein was added to HMS174(DE3) cell extract, and cell fractions were prepared as before. Each fraction was then normalized and analyzed by Western blotting using an anti-LtrA antibody. The purity of each cellular fraction was assessed by Western blotting for protein markers as in panel A. The bottom right panel shows the parallel analysis for HMS174(DE3) cells without added LtrA protein. The last lane contains the purified LtrA protein.

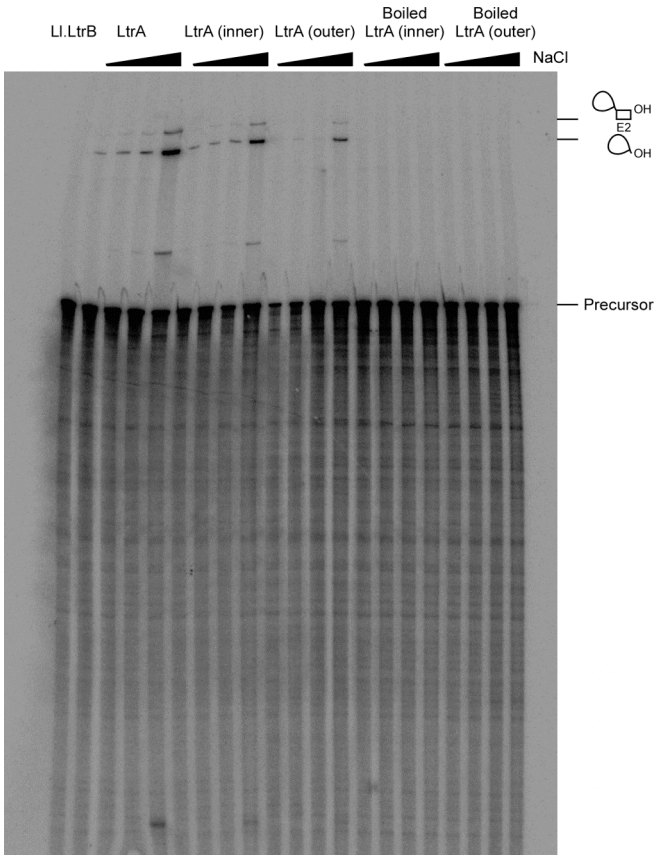



Figure 2.6: LtrA protein within membrane fractions can promote the splicing of Ll.LtrB intron precursor RNA.

100 nM of $^{32}$P-labeled Ll.LtrB precursor RNA was incubated with 20 nM LtrA protein or an equal amount of LtrA within inner or outer membrane fractions in splicing reaction buffer at 30ºC for 30 min. The reactions were terminated by phenol-chloroform-isoamyl alcohol (25:24:1) extraction, ethanol precipitation, and analyzed in a denaturing 4% polyacrylamide gel, which was dried and scanned with a phosphorimager. Lane "Ll.LtrB" contains 100 nM Ll.LtrB intron precursor RNA. Lanes "LtrA" contains LtrA protein purified from *E. coli* expressing LtrA from pIMP-lp (Saldanha et al. 1999). Lanes "LtrA (inner)" and "LtrA (outer)" are the inner or outer membrane fractions containing LtrA protein prepared from HMS174(DE3) cells expressing plac-LtrA. The boiled LtrA (inner) and LtrA (outer) are the inner or outer membrane fractions containing LtrA protein treated by heating at 95ºC for 5 min. The splicing activity of each sample was checked at different NaCl concentration (0, 50, 100 and 500 mM). The larger spliced product represents the spliced intron lariat with 3' exon attached, and the smaller spliced product is the spliced intron lariat, as indicated schematically to the right of the gel.

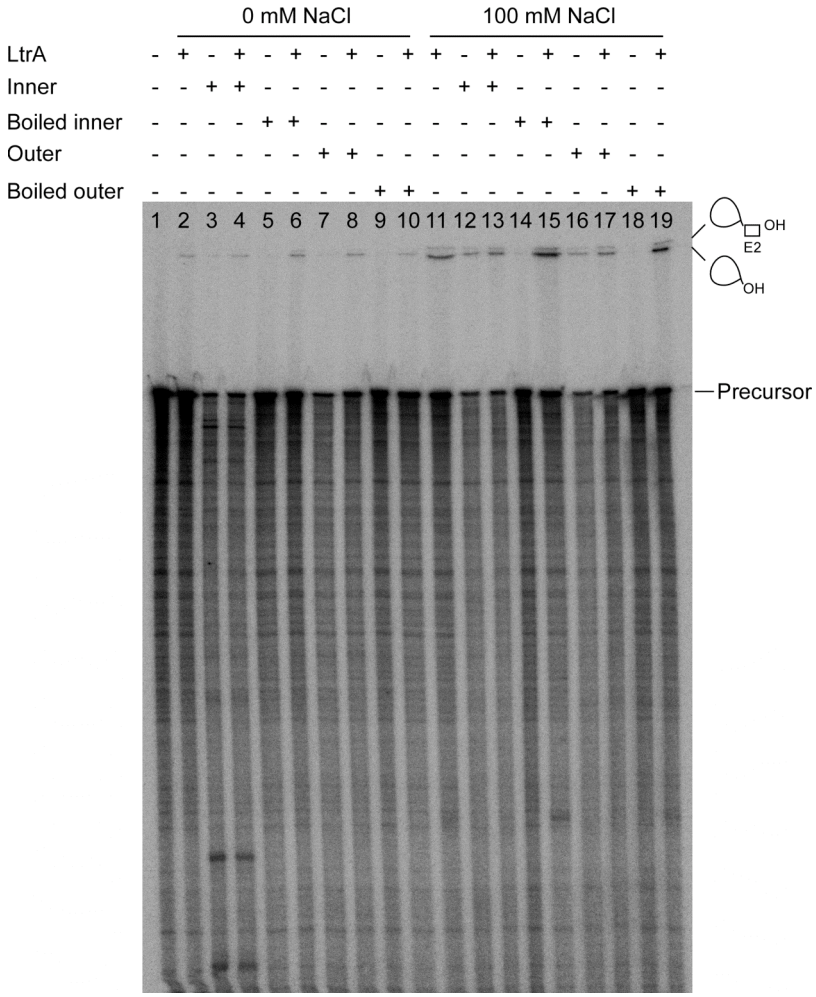



Figure 2.7: LtrA in membrane fractions can stimulate Ll.LtrB intron splicing at lower NaCl concentration than purified LtrA protein.

100 nM of $^{32}$P-labeled Ll.LtrB precursor RNA was incubated with 20 nM purified LtrA or the same amount of purified LtrA with 5 μl of inner or outer membrane fractions from HMS174(DE3) at 30ºC for 30 min. Two NaCl concentrations (0 or 100 mM NaCl) were tested. The reactions were terminated by phenol–chloroform–isoamyl alcohol (25:24:1) extraction followed by ethanol precipitation, and analyzed in a denaturing 4% polyacrylamide gel, which was dried and scanned with a phosphorimager. Boiled inner or outer indicates the membrane fraction treated by heating at 95ºC for 5 min. The splicing products lariat with exon 2 and lariat are shown schematically to the right of the gel. The addition of protein or membrane fraction is indicated at the top of gel by a "+".

# Chapter 3: EcI5, a group IIB intron with high retrohoming frequency: DNA target site recognition and use in gene targeting

Group II intron EcI5 was discovered in the *Escherichia coli* virulence plasmid pO157 (Burland et al. 1998). It is a CL/IIB1 intron, whose IEP contains an En domain (Figure 1.6B) (Burland et al. 1998; Dai and Zimmerly 2002b). The intron was found to be actively mobile in *E. coli* in a screen of candidate introns using donor and recipient plasmid vectors that were used previously for mobility assays with Ll.LtrB (Guo et al. 2000; Karberg et al. 2001; Karberg 2005). Previous studies with the group IIA intron Ll.LtrB showed that it could be retargeted to insert into different DNA target sequences by modifying the base-pairing sequences of the intron RNA (Guo et al. 2000; Mohr et al. 2000). This feature made it possible to develop Ll.LtrB into a highly efficient bacterial gene targeting vector ("targetron"), which has programmable target specificity. Thus far, no other group II intron has been shown to be similarly useful for gene targeting, and it is not clear to what extent the required combination of characteristics would be found among other group II introns. Here, I show that EcI5 has very high retrohoming frequency in *E. coli*. Its DNA target site recognition rules were determined and this information was used to develop a computer algorithm that enables EcI5 to be retargeted to insert into new sites with high frequency and specificity.

## 3.1 Construction of an EcI5-ΔORF intron and its use in genetic assays of intron mobility

Previous studies showed that the retrohoming frequency of the *L. lactis* Ll.LtrB intron expressed from a pACD-based donor plasmid could be increased as much as 40,000-fold by deleting intron ORF sequences from DIV and expressing the IEP from a position downstream of the 3' exon on the same plasmid (Guo et al. 2000). This

dramatically increased retrohoming frequency is thought to be due largely to the decreased nuclease susceptibility of the smaller ΔORF intron in *E. coli*. With this configuration, mobility frequencies for Ll.LtrB were high enough to carry out selection experiments using libraries of donor and recipient plasmids with randomized recognition elements, enabling determination of detailed DNA target site recognition rules for that intron (Guo et al. 2000; Zhong et al. 2003).

To be able to do similar experiments for EcI5, a modified intron-donor plasmid pACD2-EcI5, which expresses an EcI5-ΔORF intron and short flanking exons, with the IEP expressed from a position downstream of the 3' exon was constructed (Figure 3.1A). The EcI5-ΔORF intron was constructed by replacing intron ORF sequences (nucleotide residues 733-2265) from subdomain IVb with an MluI site (Figure 1.6A, inset upper right). The latter provides a convenient location for cloning cargo genes into the intron. The EcI5-ΔORF intron retains subdomain DIVa, which is a high-affinity binding site for the IEP in group IIA introns (Wank et al. 1999; Huang et al. 2003), as well as the residual DIVb stem, and subdomain DIVc, a small stem-loop structure found at the 3' end of DIV (see Figure 1.6A). For quantitative mobility assays, a phage T7 promoter was cloned into the MluI site in DIV. The EcI5-ΔORF intron expressed from the pACD2-EcI5 donor plasmid was confirmed to splice efficiently in *E. coli* by RT-PCR (data not shown).

For mobility assays, donor plasmid pACD2-EcI5 and the recipient plasmid pBRR3-EcI5, which contains the EcI5 target site (positions -30 to +15 from the intron-insertion site) cloned upstream of a promoterless $tet^R$ gene, were co-transformed into *E. coli* HMS174(DE3). After induction of intron-expression with IPTG, integration of the EcI5-ΔORF intron carrying the phage T7 promoter into the recipient plasmid target site activates the expression of the downstream $tet^R$ gene, enabling selection of mobility

events by plating on LB agar containing tetracycline and ampicillin. Mobility efficiencies were then calculated as the ratio of $Tet^R$ + $Amp^R$ to $Amp^R$ colonies.

As shown in Figure 3.2A, the EcI5-ΔORF intron expressed from pACD2-EcI5 at a low level without IPTG induction had a mobility efficiency of 77% in this assay, about three-times higher than a similarly configured Ll.LtrB-ΔORF intron donor plasmid assayed in parallel (24%). When induced with 100 μM IPTG, both the Ll.LtrB-ΔORF and EcI5-ΔORF introns have mobility efficiencies close to 100%. As expected, mutations in the conserved YADD motif in the RT domain of the EcI5 IEP or a C-terminal truncation that deletes both the putative D and En domains (ΔD/En) abolished EcI5 mobility without or with IPTG induction. Further, En-domain mutations (C-terminal truncation ΔEn or mutations H528A and H552A in the putative En active site) decreased EcI5 mobility frequencies without IPTG induction to < 0.61% (Figure 3.2A). Surprisingly, however, these En-domain mutants still gave high mobility when induced with 100 μM IPTG (Figure 3.2A), presumably by using alternate En-independent retrohoming mechanisms (Zhong and Lambowitz 2003). Together, these findings suggest that EcI5 uses retrohoming pathways similar to those for the Ll.LtrB intron, but with substantially higher integration frequencies, enabling the alternative En-independent pathways to operate at high efficiency.

### 3.2 OVERPRODUCTION OF ECI5 IS NOT TOXIC TO *E. COLI*

In the experiment of Figure 3.2B, it was surprising that IPTG-induction of EcI5 expression in the presence of the EcI5 recipient plasmid is ~100-fold more toxic to *E. coli* than is induction of the Ll.LtrB intron in the presence of its recipient plasmid. As a potential gene targeting tool, it is a concern that this toxicity might reflect a high frequency of ectopic integration or indiscriminate cleavage of *E. coli* chromosomal DNA

by the over-expressed intron. Additional controls, however, showed that elevated toxicity is observed only when EcI5 is expressed in the presence of its own recipient plasmid and that EcI5 expressed by itself or with the Ll.LtrB recipient plasmid is not significantly more toxic than is expression of the Ll.LtrB intron or the benign protein GFP from the same donor plasmid (Figure 3.2C). A likely explanation is that the integration frequency of EcI5 into its recipient plasmid target site is so high that multiple plasmids are targeted in each cell, leading to over-expression of the *tet*$^R$ gene, which is known to be toxic (Eckert and Beck 1989). Thus, the experiment of Figure 3.2A may substantially underestimate the mobility frequency difference between EcI5 and Ll.LtrB.

## 3.3 EcI5 INTRON DNA TARGET SITE RECOGNITION RULES

### 3.3.1 Identification of critical nucleotide residues in the distal 5'-exon and 3'-exon regions of the EcI5 DNA target site

To determine DNA target site recognition rules for the EcI5 intron, a plasmid-based Tet$^R$ selection assay was carried out using donor and recipient plasmid libraries in which different recognition elements were randomized. As an overview for these experiments, Figure 3.3 (top) shows the DNA target sequence for EcI5 and its predicted base-pairing interactions with the intron RNA. The EBS2, EBS1, and EBS3 sequences, each located in a different region of DI, potentially base pair to DNA target site sequences IBS2, IBS1, and IBS3, respectively, spanning positions -13 to +1 from the intron-insertion site. By analogy with the Ll.LtrB intron, the IEP is expected to recognize DNA target site positions both upstream and downstream of those that base pair with the intron RNA.

To identify nucleotide residues potentially recognized by the IEP, Michael Karberg carried out a selection experiment in which the wild-type EcI5-ΔORF intron

integrates into recipient plasmid target sites with randomized nucleotide residues from positions -35 to -14 and +2 to +20 of the target sequence. After plating on LB agar containing tetracycline and ampicillin, plasmids containing integrated introns were isolated by a mini-prep procedure, and the sequences of active target sites were determined by sequencing the 5'- and 3'-integration junctions. In the experiment of Figure 3.3, target sequences were obtained from 101 active target sites selected in the experiment and 108 plasmids from the initial pool to correct for nucleotide frequency biases, as described in Methods. Figure 3.3 (bottom) summarizes the results in WebLogo format, along with nucleotide frequencies at each randomized position in both the selected target sites and the initial pool.

In the distal 5'-exon region, the most strongly conserved nucleotide residues in the target sequence were C-18, C-17, A-15, and A-14, with A-14 found in 100% of the active target sites, and in the 3' exon, the most strongly conserved nucleotide residue was T+5, which was found in 80% of active target sites. Several other 3'-exon positions in the EcI5 target site showed some selection for specific nucleotide residues (+2, +3, +4, +6, and +10). In agreement with the selection experiment, I found that single nucleotide substitutions at A-14, which was found in 100% of active target sites, strongly decreased mobility frequencies (A-14C, A-14G, and A-14T, 1. 0, 1.4, and 1.6% of wild type, respectively, average of two assays done as in Figure 3.2 without IPTG induction). As expected, single nucleotide substitutions at other key positions identified in the selections also inhibited mobility but less severely than mutations at A-14 (C-18T, A-15C, and T+5C, 8%, 2%, and 20% of wild type, respectively, without IPTG induction).

In addition to identifying nucleotide residues potentially recognized by the IEP, the selection data in Figure 3.3 suggest that the EBS3/IBS3 interaction is limited to

position +1, because positions +2 and +3 do not show strong selection for nucleotide residues that could base pair with U-residues in the intron RNA to extend this interaction. DNA target site position +2 does show weak selection for a non-wild-type A-residue that could base pair with the U-residue at EBS3 +2, but also shows similar weak selection for a non-wild-type T-residue that cannot base pair with this U-residue. These weak selections may reflect that a weak A-T or T-A DNA base pair at target site position +2 helps accommodate the EBS3/IBS3 interaction between the intron RNA and DNA target site at position +1.

The selected target sites have GC contents ranging from 18-77% from position -35 to -14 and 21-74% from position +2 to +20, indicating that EcI5 can integrate into DNA target sites having a wide range of melting temperatures (not shown). Analysis of 30 selected target sites using DNA Mfold (http://mfold.bioinfo.rpi.edu/cgi-bin/dna-form1.cgi) did not reveal any conserved stem-loop structures in the EcI5 DNA target site. Using a chi-square test with probability level P = 0.001 and nine degrees of freedom, two potential co-variations between nucleotide residues in the DNA target sequence (positions -24 and -18; $chi^2$ = 30.3 and positions +9 and +11; $chi^2$ = 28.8) were found. However, both are based on dinucleotides with low (< 10) observed counts, and their significance in this relatively small dataset is unclear.

### 3.3.2 Rules for the EBS1/IBS1 pairing

The predicted EBS1/IBS1 interaction between wild-type EcI5 and its DNA target site encompasses positions -5 to -1 from the intron-insertion site (Figure 3.4). To determine rules for base pairing between EBS1 and IBS1, the positions -7 to -1 in the EBS1 stem-loop in the intron and the corresponding positions in the IBS1 region of the DNA target site were both randomized (Figure 3.4A). Because the EBS1/IBS1 pairing is

also required for RNA splicing, IBS1 positions -7 to -1 in the 5' exon of the donor plasmid were also randomized to provide complementary nucleotide combinations in the pool (Guo et al. 2000). After selection, the 5'-integration junctions were amplified from the selected $Tet^R$ + $Amp^R$ colonies by colony PCR, and the PCR products were sequenced to identify active EBS1/IBS1 combinations. Sequences for 87 independent intron-integration events, along with 94 unselected donor plasmids and 95 unselected recipient plasmids to correct for nucleotide frequency biases in the initial pools, were obtained. Figure 3.4B summarizes nucleotide frequencies at each of the randomized positions, with the results depicted in WebLogo format above, while Figures 3.4C and D and Table 3.1 summarize base-pair frequencies between intron RNA and DNA target site positions.

The results show selection for base pairing between the intron RNA and DNA target site from positions -6 to -1, with no selection for base pairing at position -7 (Figure 3.4C; Table 3.1). The selection for base pairing at position -6 was unexpected because wild-type EcI5 cannot form a canonical Watson-Crick or wobble base pair with its DNA target site at this position. Ninety five percent of the selected intron/target site combinations have four or more base pairs in EBS1/IBS1, compared to only 21% for randomly paired introns and target-site combinations from the original pools (Figure 3.4C).

In the intron RNA, EBS1 positions -6 to -1, which base pair with the DNA target site, show some selection for wild-type nucleotide residues, with the strongest selection for C-6, G-5, and U-2 (Figure 3.4B). There is also weaker selection for complementary nucleotide residues in IBS1, leading to some preferences for specific base pairs between the intron RNA and DNA target site (*e.g.*, C-G at position -6, G-C or G•T at position -5 and U-A at position -2; Table 3.1; also observed when data are corrected for nucleotide

frequency biases; data not shown). The most likely explanation is that the selection for specific EBS1 nucleotide residues reflects constraints on the structure of the EBS1 RNA stem-loop, which in turn leads to selection for complementary nucleotide residues in IBS1. However, the weak selection for some IBS1 nucleotide residues also reflects constraints on DNA structure or partial recognition of IBS1 sequences by the IEP. It is also notable that some EBS1/IBS1 positions show selection for non-wild-type base pairs (*e.g.,* U-A or C-G instead of U•G at position -4 and C-G instead of U-A at position -3; Table 3.1). Notably, EBS1 position -7, which does not base pair to the DNA target site, nevertheless shows strong selection for the wild-type U residue, almost certainly reflecting a constraint on the structure of the RNA stem-loop, as posited for the other EBS1 nucleotide residues (Figure 3.4B).

### 3.3.3 Rules for the EBS2/IBS2 pairing

Figure 3.5 shows the results of similar selection experiment for positions -13 to -7 potentially involved in the EBS2/IBS2 interaction. In this experiment, I obtained sequences of 102 independent integration events, along with 99 unselected donor plasmids and 103 unselected recipient plasmids to correct for nucleotide frequency biases in the pools. The data show strong selection for RNA/DNA base pairing between positions -13 and -9, with selection against RNA/DNA base pairing at position -8, and no selection for or against RNA/DNA base pairing at position -7 (Figure 3.5C, D).

Notably, the selection for base pairing in EBS2/IBS2 appears stronger than in EBS1/IBS1, with each position between -13 and -9 base paired in ≥ 95% of the selected combinations and positions -10 and -9 base paired in 100% and 99% of the selected combinations, respectively (Figure 3.5C). Additionally, 91% of the selected combinations have all five base pairs between positions -13 and -9, while randomly chosen intron-

target site combinations from the original pool typically have three or fewer base pairs between these positions (Figure 3.5D).

Within the region of EBS2 involved in base pairing (positions -13 to -9), positions -11 to -9 show moderate to high selection for specific nucleotide residues (Figure 3.5B). As in the case of EBS1, the most likely possibility is that the selection for these nucleotide residues reflects constraints on RNA structure, and these constraints in turn result in preferences for specific RNA-DNA base pairs at these positions (Table 3.1). Positions -11 and -10 show particularly strong selection for the wild-type C-G pairs, but some selection for C-G base pairs is also seen at positions -12 and -9, where the wild-type base pair is in both cases A-T. In contrast to the EBS1/IBS1 interaction, when the data are corrected for nucleotide frequency biases in the initial pools, the wild-type EBS2/IBS2 base pair appears to be preferred at each position (data not shown).

Intron RNA positions -8 and -7, which are not involved in base pairing with the DNA target site, nevertheless show strong selection for the wild-type A- and U-residues, respectively, again likely reflecting constraints on intron RNA structure (Figure 3.5C). Position -7 in the DNA target site shows no selection for a specific nucleotide residue, in agreement with the results in the EBS1/IBS1 selection above, while position -8 in the DNA target site shows some selection for G or A residues, possibly reflecting prohibition against base pairing with the strongly conserved A-residue at position -8 in the intron RNA.

### 3.3.4 Rules for the EBS3/IBS3 pairing

The EBS3/IBS3 interaction in EcI5 is predicted to involve a single base pair at position +1 (Figure 3.6). To confirm this prediction and assess nucleotide preferences for the EBS3/IBS3 interaction, a final selection experiment was carried out with a random

nucleotide in position +1 both in IBS3 of the recipient plasmid and EBS3 of the donor plasmid (Figure 3.6A). The results show selection for RNA/DNA base pairing, with 71% of the selected target sites having either a Watson-Crick or wobble G•T or U•G pair at position +1, compared to only 22% in the pool (Figure 3.6B). Notably, although all four Watson-Crick combinations and some mismatches (RNA/DNA C/A and G/A) are reasonably well represented, some nucleotide mismatches (A/C, U/C, A/G, G/G, C/T, and U/T) were not found, even though a relatively large number of events were analyzed for a selection involving a single base pair (Table 3.1). These findings raise the possibility that the EBS3/IBS3 interaction may be particularly sensitive to some mismatches.

For this selection, position +1 in the 3' exon of the donor plasmid was not randomized, and the donor plasmid retained the wild-type C-residue at this position. For the Ll.LtrB intron, the analogous δ-δ' interaction between the intron and 3' exon in the precursor RNA may contribute to the efficiency of RNA splicing, but generally has only a small effect on the overall intron-integration efficiency (Perutka et al. 2004). Our results suggest that this is also the case for the EBS3/IBS3 pairing in EcI5, with about half of the selected introns, those with A, C, or T residues at EBS3 +1, unable to form an EBS3/IBS3 base pair in the precursor RNA. Nevertheless, selection for an EBS3/IBS3 base pair in the precursor RNA could contribute along with constraints on RNA structure to the observed bias for introns with the wild-type G-residue at EBS3 +1.

### 3.4 A COMPUTER ALGORITHM FOR IDENTIFYING EcI5-INSERTION SITES AND RETARGETING EcI5

The DNA target site recognition rules revealed by the selection experiments were further used to develop a computer algorithm that identifies potential EcI5 insertion sites and designs PCR primers for modification of the intron's EBS sequences to insert into those sites. As done for the Ll.LtrB intron (Perutka et al. 2004), a simple probabilistic

model (zero-order Markov model), which assumes that the likelihood of occurrence of a nucleotide residue in the target sequence is independent of other residues in this sequence, was used. Model parameters are derived from the observed frequency of residues in a training set of trusted examples. Thus, the performance of the model critically depends on which target site positions are included in the calculation. Because the model is based on a limited amount of data, overtraining can occur, resulting in a model that is able to classify sequences from the training set correctly, but unable to make accurate predictions for sequences not drawn from the original training set. The more parameters that are included in the model, the greater the risk of overtraining.

By a procedure described below, a minimal set of DNA target site positions sufficient for discriminating efficient target sites from inefficient ones includes: -26 to -14, -8, -6, and +2 to +10, with A-14, which is present in 100% of the selected target sites, treated as a fixed position in the model. Positions -26 to -14 and +2 to +10 are the regions presumably recognized by the IEP. Positions -6 and -8, which are not thought to be recognized by the IEP, nevertheless have moderately high information content in the EBS1/IBS1 selection of Figure 3.3 and the EBS2/IBS2 selection of Figure 3.4, respectively. For position -6, this information content reflects selection for a nucleotide residue that extends the RNA/DNA base-pairing interaction, while for position -8, it reflects selection against a nucleotide residue that extends the RNA/DNA base-pairing interaction.

The algorithm scores potential DNA target sites across a 36-bp sliding window with 1-bp increments by calculating a log-odds score *S*, which compares the probability $P(seq|M)$ that the target sequence *seq* was generated by the model *M* to the probability $P(seq)$ that the sequence was generated by chance (null model). $P(seq|M)$ and $P(seq)$ are

calculated as the products of individual probabilities $P(n_p|M)$ and $P(n_p)$ that a nucleotide $n$ at position $p$ in the target sequence was generated by the model and by chance, respectively, with the latter calculated from the frequencies $f_{n_p(selected)}$ and $f_{n_p(pool)}$ for a nucleotide residue in selected target sites and the initial pool, respectively, according to the equation:

$$S = \log_2 \frac{P(seq \mid M)}{P(seq)} = \log_2 \frac{\prod_p P(n_p \mid M)}{\prod_p P(n_p)} = \sum_p \log_2 \frac{f_{n_p(selected)}}{f_{n_p(pool)}}$$

A positive score indicates that the model predicts the potential EcI5 target sequence better than the null model (see Perutka et al. 2004 for additional description).

### 3.5 RETARGETING OF ECI5 TO INSERT INTO SITES IN THE *E. COLI LACZ* GENE

To test the performance of different models, potential target sites in the *E. coli lacZ* gene were identified by different algorithms. The retargeted EcI5 was made by modifying its EBS1, 2, and 3 sequences to form Watson-Crick base pairs at the positions recognized by base pairing of the intron RNA (IBS2 -13 to -9; IBS1 -6 or -5 to -1; and IBS3, +1). IBS2, IBS1, and IBS3 in the 5'- and 3'-exons of the donor plasmid were also modified to be complementary to the retargeted EBS2, EBS1, and EBS3 sequences for efficient RNA splicing.

Figure 3.7 shows potential target sequences in the *lacZ* gene and their base-pairing interactions with the intron RNA, along with their log-odds scores computed by the algorithm for the model described in the preceding section. To facilitate retargeting, four different donor plasmids, pACD3-EcI5A, C, G and T were constructed. These constructs lack the T7 promoter in DIV and have the indicated nucleotide residues at EBS3 along with the complementary nucleotide residue at IBS3 for maximally efficient

RNA splicing. For retargeting, the appropriate donor plasmid whose EBS3 nucleotide residue is complementary to IBS3 in the DNA target site was used to introduce the required modifications into the donor plasmid's EBS1, EBS2, IBS1 and IBS2 sequences by PCR, as diagrammed in Figure 3.8A.

The donor plasmids containing the retargeted introns were transformed into *E. coli* HMS174(DE3). After induction of intron expression with 100 $\mu$M IPTG for 3 h at 37ºC, disruption of the *lacZ* gene was scored by plating on LB agar containing 5-bromo-4-chloro-3-indolyl-D-galactopyranoside (X-gal), where $LacZ^+$ colonies are blue and $LacZ^-$ colonies are white. For each retargeted intron, 100 to 1,000 colonies, depending on the targeting frequency, were screened and confirmed for integration at the desired site by colony PCR and sequencing both the 5'- and 3'-integration junctions for at least 12 white colonies (Figure 3.8B).

Comparison of the experimentally determined targeting frequencies with the log-odds scores computed by the algorithm for the model described in the preceding section showed that all efficient target sites (targeting frequencies 30-98%) have log-odds score higher than 8.2, while all less efficient target sites (targeting frequencies 0 to 4.0%) have log-odds scores less than 8.2 (Figure 3.7). More complicated models that included additional DNA target site positions or frequencies of EBS/IBS base pairs from the selection experiments did not give a similar clean separation of efficient and inefficient target sites. This situation likely reflects that the currently available selection data are too limited to accurately weight the effects of substituting different combinations of nucleotide residues in the EBS sequences of the intron RNA. The model will need to be improved as more data becomes available. As for Ll.LtrB targetrons, some retargeted

EcI5 introns with lower log-odds scores still gave experimentally useful targeting frequencies.

All of the introns in Figure 3.7 were tested with a 5-bp EBS1/IBS1 interaction like that of the wild-type EcI5 with its DNA target site. Because the selection experiment of Figure 3.4 showed the potential for a sixth base pair at EBS1/IBS1 position -6, the insertion frequencies of a number of retargeted introns with five or six base-pair EBS1/IBS1 DNA target site interactions were compared. In four cases (1806s, 178a, 1790a, 187s), the additional EBS1/IBS1 base pair gave the same or increased targeting frequencies, while in four other cases (912s, 1257s, 326a, 1878s), it decreased the targeting frequency significantly (Figure 3.7). These findings may reflect preferences for specific nucleotide residues at position -6 in the context of different EBS1 sequences. Until these effects are understood better, it seems preferable for retargeting EcI5 to leave the wild-type nucleotide residue at EBS1 position -6 and use the five base-pair EBS1/IBS1 interaction like that of wild-type EcI5 with its DNA target site.

The EcI5 intron was also tested by expressing it from a broad-host range vector pBL1, which can be used in diverse Gram-negative bacteria without introducing a gene encoding phage T7 RNA polymerase (Yao and Lambowitz 2007). This plasmid expresses targetrons by using an m-toluic acid-inducible promoter ($P_m$) recognized by the host RNA polymerase. For EcI5 targetron LacZ1806s, pBL1-EcI5 gave a lacZ targeting frequency of only 4% compared to 98-99% for the same targetron expressed from pACD3-EcI5A. In previous work, pBL1 gave high insertion frequencies with several Ll.LtrB targetrons in different bacteria and appeared to be as efficient as pACD3 for expressing Ll.LtrB targetron LacZ635s in *E. coli* HMS174(DE3) (Yao and Lambowitz 2007). However, additional tests comparing other Ll.LtrB LacZ targetrons in *E. coli* HMS174(DE3)

showed that some targetrons function as efficiently when expressed from pBL1 as pACD3, while others function less efficiently (Yao and Lambowitz, unpublished results). A possible explanation is that targetrons with lower splicing efficiencies due to different EBS/IBS interactions in the precursor RNA benefit from more efficient expression by T7 RNA polymerase.

For four of the retargeted EcI5 introns expressed from pACD3-EcI5 in the experiment of Figures 3.7 and 3.8, genomic DNA from the disruptants was analyzed by Southern hybridization to confirm site-specific insertion. For three disruptants, 163s, 1806s and 1790a, the Southern blots hybridized with an intron probe showed a band of the size expected for site-specific insertion into *lacZ* gene and no non-specific insertions (Figure 3.8C; the extra band for 1790a is due to residual donor plasmid). In the remaining case, 1257s, the disruptants obtained after induction with 100 μM IPTG for 3 h showed insertion at a second site in addition to the expected site in *lacZ* (not shown), but this ectopic targeting was not evident when the induction was done with a lower amount of IPTG for a shorter time (50 μM IPTG, 1 h; Figure 3.8C). Together, the above findings show that EcI5, like Ll.LtrB, can be retargeted to insert efficiently and specifically into desired chromosomal DNA targets.

### 3.6 DISCUSSION

The group II intron EcI5, a subclass CL/IIB1 intron discovered in an *E. coli* virulence plasmid (Burland et al. 1998), is highly active in retrohoming in *E. coli*, and was adapted for use in gene targeting. Like the well-studied group IIA introns *S. cerevisiae* aI1 and aI2 and *L. lactis* Ll.LtrB, EcI5 encodes a protein with conserved RT, X, D, and En domains. As expected for the retrohoming mechanism used by these group II introns, EcI5 retrohoming is abolished by mutations that inhibit IEP expression or the

ribozyme activity of the intron RNA, as well as by the mutation YADD→YAAA in the RT active site. Further, as shown for the Ll.LtrB intron (Zhong and Lambowitz 2003b), En-domain mutations strongly inhibit EcI5 retrohoming, but leave residual retrohoming by En-independent mechanisms. It is a measure of the high activity of EcI5 that such En-independent retrohoming can occur at almost 100% efficiency when intron-expression is induced with IPTG.

Both full-length EcI5 and a streamlined EcI5-ΔORF derivative with the IEP expressed from the same donor plasmid have higher mobility frequencies than do the corresponding Ll.LtrB intron constructs. These higher mobility frequencies could reflect more efficient production or higher stability of the intron RNA or IEP, more active RNPs, or the ability to efficiently use both En-dependent and En-independent retrohoming mechanisms. Biochemical analysis will be needed to determine whether EcI5 RNPs have inherently higher DNA integration efficiency than do Ll.LtrB RNPs. As for Ll.LtrB (Guo et al. 2000), the deletion of ORF sequences from DIV enables EcI5 to retrohome at near 100% efficiency, presumably by decreasing the nuclease susceptibility of the intron RNA, which appears to be a major factor limiting the mobility of group II introns in *E. coli* (Guo et al. 2000; Smith et al. 2005).

EcI5 was found in virulence plasmid pO157 in *E. coli* strain O157:H7 (Burland et al. 1998; Dai and Zimmerly 2002b). The intron is inserted within a non-coding region of the plasmid, but in the same orientation as most other plasmid genes. Southern hybridizations and PCR analysis showed that EcI5 is also inserted at the same site in DNAs isolated from 19 of 72 ECOR strains, although always in fragmented form (Dai and Zimmerly 2002b). The computer algorithm developed here for identifying EcI5 target sites found no efficient target sites for wild-type EcI5 intron (log-odds score > 8.2

and Watson-Crick or wobble base pairs at all EBS/IBS positions) in sequenced *E. coli* O157 genomes (EC4115, EDL933, and Sakai), and only one such target site in the non-essential gene *cusS* in the genome of *E. coli* K12 MG1655, which does not carry the plasmid. Thus, the spread of EcI5 to the *E. coli* chromosome may be limited by a combination of poor expression from a non-coding region of pO157 and a paucity of efficient insertion sites in the genome.

Like other mobile group II introns, EcI5 recognizes DNA target sequences by using both the IEP and base pairing of the intron RNA. As expected for a class IIB intron, the base-pairing interactions involve EBS1 and EBS2, which base pair to the 5' exon, and EBS3, which base pairs to the 3' exon (Costa et al. 2000; Jiménez-Zurdo et al. 2003). Detailed analysis by selection experiments using randomized sequence libraries showed that the EBS1/IBS1 interaction can extend for six base pairs rather than five base pairs found for the interaction of the wild-type intron with its DNA target site. The selection experiments also suggest relatively strong constraints on the EBS2 sequence, likely reflecting the deleterious effect of nucleotide substitutions in the EBS2 loop on intron RNA structure. The constraints on the EBS2 sequence appear to be greater in EcI5 than in Ll.LtrB, perhaps reflecting that EBS2 is located in a junctional loop in the former and a stem-loop in the latter, a difference between IIB and IIA introns.

As for other mobile group II introns, the EcI5 IEP appears to recognize sequences in both the distal 5'-exon and 3'-exon regions of the DNA target site. In the distal 5'-exon region, the number of nucleotide residues critical for EcI5 recognition is similar to that for Ll.LtrB and other mobile group II introns, but their identity and location are different. Thus, for EcI5, the most critical distal 5'-exon nucleotide residues are C-18, C-17, A-15, and A-14, with A-14 found in 100% of active target sites (Figure 3.3), while

for Ll.LtrB, the most critical nucleotide residues in this region are T-23, G-21, A-20, T-19, and G-15, with none as stringently required as A-14 for EcI5 (Singh and Lambowitz 2001; Perutka et al. 2004). In the 3' exon, EcI5 is similar to Ll.LtrB in that the only critical nucleotide residue is T+5. In Ll.LtrB, T+5 is required for second-strand cleavage by the En domain, but not for the initial DNA target site recognition or reverse splicing (Mohr et al. 2000). If that is also true for EcI5, retrohoming of EcI5 to sites lacking T+5 may occur by En-independent pathways. Although difficult to compare due to analysis by different methods, the target sequences for group II IEPs studied thus far have almost no critical nucleotide residue in common, and this divergence in IEP recognition sequences is seen even for very closely related introns, such as *S. cerevisiae* aI1 and aI2 (Guo et al. 1997; Yang et al. 1998; Singh and Lambowitz 2001; Jiménez-Zurdo et al. 2003). These findings suggest that IEP recognition sequences can evolve rapidly to adapt the intron to retrohome to new sites (Singh and Lambowitz 2001).

The selection experiments enable us to calculate values for the total information content of nucleotide residues putatively recognized by the IEPs of different introns. For EcI5, this value calculated for positions -26 to -14 and +3 to +10 is 6.96 bits, while for Ll.LtrB the value calculated for the same nucleotide positions is 5.21 bits (using nucleotide frequency data of Zhong et al. 2003). As both EcI5 and Ll.LtrB are similar in potentially forming 11 or 12 base pairs with their DNA target sites, these findings suggest that target site recognition for EcI5 may be somewhat more stringent than for Ll.LtrB. Attesting to their very high specificity, over-expression of neither EcI5 nor Ll.LtrB is toxic to *E. coli*, as would be expected if these introns could indiscriminately cleave or insert into non-target sequences. The high target specificity of mobile group II introns reflects that 11-15 nts of the DNA target sequence are recognized by base pairing of the

intron RNA and that base-pairing mismatches are expected to strongly affect the $k_{cat}$ for reverse splicing in addition to $K_m$, thereby limiting off-target insertions (Xiang et al. 1998).

As a gene-targeting vector, the high retrohoming efficiency and novel target specificity of EcI5 expand the number of genomic target sites that can be targeted efficiently by group II introns. Analysis of the *E. coli* K12 genome for potential EcI5 targeting sites using the computer algorithm developed here revealed 14,938 such sites on either DNA strand with log-odds scores > 8.2, an average of one highly ranked target site per 621 nucleotide residues. In the *E. coli LacZ* gene (3-kb), the Ll.LtrB intron has five potential targeting sites with log-odds scores > 8.2, of which two have been validated with targeting efficiencies > 10% (Yao and Lambowitz 2007; Zhao et al. 2008), while EcI5 has six completely different targeting sites with log-odds scores > 8.2 of which five have been validated (targeting efficiencies 30-98%). Including less efficient introns whose targeting sites have lower log-odds scores, a total of ten EcI5 target sites were validated in *lacZ* (Figure 3.7). To date, 219 full-length ORF-containing group II introns have been identified in different bacteria, and their insertion sites suggest a wide variety of IEP recognition sequences (Simon et al. 2008). Assuming that a significant fraction of these bacterial group II introns are mobile and can be adapted for gene targeting as shown here for EcI5, the diversity of homologous IEPs with novel target specificity should greatly expand the number of target sites accessible to group II introns without resort to protein engineering.

Finally, EcI5 has a combination of characteristics that make it well-suited for gene targeting applications. These characteristics include high retrohoming efficiency, novel IEP recognition specificity, relatively long EBS/IBS interactions between the

intron RNA and DNA target, sufficient flexibility in its EBS sequences to retarget the EBS/IBS interactions, and a high degree of target specificity. The required combination of characteristics for gene targeting is most likely to be found among group IIA and IIB introns, whose IEPs have an En domain, enabling them to retrohome efficiently to both DNA strands without relying on nascent strands at DNA replication forks to prime reverse transcription. Group IIC introns, which recognize DNA stem-loop structures, have short EBS1/IBS1 interactions, and encode proteins lacking an En domain, are likely to have insufficient flexibility or specificity for gene targeting. The procedures described here for constructing a highly active EcI5-ΔORF derivative and rapidly determining targeting rules should be generally applicable to other group II introns, which when similarly developed will provide a wide range of different IEP target specificities and potentially other useful characteristics.

## 3.7 METHODS

### 3.7.1 Bacterial strains and growth conditions

*E. coli* HMS174(DE3) *recA1* (Novagen) was used for intron-mobility experiments, and DH5α and DH10B (Invitrogen) were used for cloning. Unless specified otherwise, cells were grown in Luria-Bertani (LB) medium, and antibiotics were added at the following concentrations: ampicillin, 100 μg/ml; chloramphenicol, 25 μg/ml; tetracycline, 25 μg/ml.

### 3.7.2 Recombinant plasmids

Previously described donor plasmids for the Ll.LtrB intron were pACD-LtrB, which expresses the full-length intron (Guo et al. 2000) and pACD2X, which expresses a 0.906-kb Ll.LtrB-ΔORF intron with a phage T7 promoter inserted in DIV (San Filippo

and Lambowitz 2002). Both are pACYC184-based plasmids that carry a *cam*$^R$ marker on the vector backbone and use a T7lac promoter to express the intron and short flanking exon sequences. In pACD2X, the ORF encoding the IEP (LtrA protein) is cloned downstream of the 3' exon.

Donor plasmids for EcI5 are derivatives of pACD2X. pACDF-EcI5 contains the full-length EcI5 intron and flanking exon sequences swapped for the Ll.LtrB and LtrA segments of pACD2X. It was constructed by PCR amplifying the EcI5 segment of *E. coli* virulence plasmid pO157 (Burland et al. 1998), using primers EcI5-5'exon and EcI5-3'exon, which append XbaI and PstI sites, respectively (Table 3.2 for primer sequences). The PCR products were 5' phosphorylated with phage T4 polynucleotide kinase (New England Biolabs), blunt-ended with T4 DNA polymerase (Invitrogen), and cloned between the blunt-ended HindIII and XhoI sites of pACD2X. Mutant derivatives of pACDF-EcI5 with a deletion of intron DV (intron positions 2332-2365; ΔDV) or a point mutation in the initiation codon of the intron ORF (ATG→ATT; G611T) were constructed via PCR with primers containing the mutations.

pACD3-EcI5 contains a 906-nt EcI5-ΔORF intron and flanking exons, with the IEP cloned downstream of the 3' exon. To construct this plasmid, 5'- and 3'-segments of EcI5 were amplified separately by PCR of plasmid pO157 DNA (see above), using primers EcI5-5'exon + EcI5-P1 and EcI5-P3 + EcI5-3'exon, respectively (Table 3.2). The two PCR products were then mixed and amplified with the outside primers EcI5-5'exon and EcI5-3'exon to generate a 0.961-kb PCR product containing the EcI5-ΔORF intron and flanking exons with an MluI site inserted at the site of the ORF deletion in DIV. The PCR product was blunt-ended with T4 DNA polymerase (Invitrogen) and swapped for the Ll.LtrB-ΔORF intron between the XbaI and PstI sites of pACD2X (see

above). The LtrA ORF encoded downstream of the intron in pACD2X was then deleted by a single-primer PCR using primer E2-3'ORF (Table 3.2), which simultaneously inserts a SwaI site. Finally, the ORF encoding the EcI5 IEP was amplified from pACDF-EcI5 by PCR using the primers EcI5-5'ORF and EcI5-3'ORF (Table 3.2) to generate a 1780-bp product, which was blunt-ended with T4 DNA polymerase and cloned between the SwaI and SnaBI sites of the vector. Primer EcI5-5'ORF inserts the Shine-Dalgarno sequence of expression vector pET11c (New England BioLabs) upstream of the EcI5 IEP ORF.

pACD2-EcI5 was derived from pACD3-EcI5 by cloning a T7 promoter sequence (annealed 5'-phoshorylated oligonucleotides T7MLU1-T and T7MLU1-B; Table 3.2) into the MluI site in DIV. Derivatives of pACD2-EcI5 with EcI5 ORF mutations YADD→YAAA (ORF nucleotide residues 838-843 changed from GATGAT to GCGGCG); H528A (ORF nucleotide residues 1582-1584 changed from CAT to GCT); and H552A (ORF nucleotide residues 1654-1656 changed from CAC to GCC) were constructed by site-directed PCR mutagenesis with primers that contain the mutant sequence, using a QuikChange Site-Directed Mutagenesis kit (Stratagene). C-terminal truncations ΔD/En (deleted amino acid residues 429-574), and ΔEn (deleted amino acid residues 518-574) were constructed by PCR with primer DelEnS, which anneals to an upstream ORF sequence containing a BglII site, and downstream primer DelD/EnA or DelEnA, respectively, which anneal at the site of the truncation and add a TGA stop codon and a SpeI site (Table 3.2). The PCR products were then digested with BglII and SpeI and swapped for the corresponding fragment of pACD2-EcI5.

pACD3-EcI5A, C, G, or T are derivatives of pACD3-EcI5 in which EBS3 is either A, C, G or T and IBS3 is the complementary nucleotide residue. The plasmids

were constructed by site-directed PCR mutagenesis (QuikChange Site-Directed Mutagenesis kit; Stratagene), using pACD3-EcI5 as the template, with primers EBS3N and IBS3N (Table 3.2), where N is the desired EBS3 or IBS3 nucleotide residue.

Donor plasmid pBL1 is a derivative of the broad-host-range vector pJB866 (Blatny et al. 1997), which uses an m-toluic acid-inducible promoter $P_m$ to express a cassette consisting of the Ll.LtrB-ΔORF intron with flanking exons followed by the IEP (Yao and Lambowitz 2007). A parallel construct expressing EcI5 targetron LacZ1806s was constructed by PCR amplifying the EcI5-ΔORF intron, flanking exon sequences, and EcI5 IEP from pACD3-EcI5A containing the retargeted intron (see below), using primers EcI5 5' and EcI5 3' (Table 3.2), then digesting the PCR product with HindIII and XhoI, and cloning it between the corresponding sites of pJB866.

Recipient plasmid pBRR3-ltrB contains the Ll.LtrB intron-insertion site (ligated exon 1 and 2 sequences of the *ltrB* gene from position -30 upstream to +15 downstream of the intron-insertion site) cloned upstream of a promoterless $tet^R$ gene in an $Amp^R$ pBR322-based vector (Guo et al. 2000; Karberg et al. 2001). pBRR-EcI5 and pBRR3-EcI5 contain different lengths of the EcI5 insertion site (*i.e.*, ligated 5'- and 3'-exon sequences flanking EcI5 in virulence plasmid pO157) cloned in place of the Ll.LtrB insertion site in pBRR3-ltrB. The ligated-exon sequences in pBRR-EcI5 extend from positions -30 to +5 from the intron-insertion site, and those in pBRR3-EcI5 extend from positions -30 to +15. The plasmids were constructed by annealing two 5'-phosphorylated oligonucleotides, which contain the EcI5 target sequences and append AatII and EcoRI sites, and then swapping the annealed oligonucleotides for the AatII and EcoRI fragment of pBRR3-ltrB.

All constructs were sequenced to confirm the expected modifications and lack of adventitious mutations in regions subjected to PCR.

### 3.7.3 Donor and recipient plasmid libraries

pACD2-EcI5 donor plasmid libraries used in selection experiments contain randomized nucleotide residues at EBS1 positions -7 to -1, EBS2 positions -13 to -7, or EBS3 position +1. The EBS1 and EBS2 libraries also contain randomized nucleotide residues at the corresponding IBS1 and IBS2 positions in the 5' exon to provide complementary sequences that can base pair with the randomized EBS1 and EBS2 sequences for RNA splicing. The libraries were constructed by a two-step PCR of pACD2-EcI5 with primers that introduce the random nucleotide residues. In the first step, two parallel PCRs were done using the following primers: EBS1 library, primers EcI5IBS1-N + EcI5BASWT and EcI5EBS1-N + EcI5AVA2AS; EBS2 library, primers EcI5IBS2-N + EcI5EBS2-N, and EcI5EBS1WT + EcI5AVA2AS; and EBS3 library, primers pACD229s + EcI5 347a and EcI5EBS3N + EcI5AVA2AS (Table 3.2). In the second step, the gel-purified PCR products from the first step were used as templates in a second PCR with the outside primers (underlined), which append XbaI and AvaII sites. The products were digested with XbaI and AvaII and swapped for the corresponding fragment of a pACD2-EcI5 intermediate plasmid that contained a stuffer 33-bp oligonucleotide inserted between the XbaI and AvaII sites. The use of this intermediate eliminates the possibility of library contamination by undigested or self-ligated wild-type donor plasmids.

Recipient plasmid libraries containing random nucleotide residues at different target site positions were constructed by PCR using primers that introduce random bases at the desired positions and append AatII and EcoRI sites, enabling the segment to be

swapped for the corresponding segment of pBRR3-ltrB. Libraries and the primers used to construct them were: -35 to -14/+2 to +20 library, EcI5-5'N22 + EcI5-3'N19; IBS1 library, EcI5 IBS1A + EcI5 IBS1S; IBS2 library, EcI5 IBS2A + EcI5 IBS2S; and IBS3 library, EcI5 IBS3A + EcI5IBS3S (Table 3.2).

After construction, the libraries were electroporated into *E. coli* DH10B or DH5α, and the cells were grown overnight at 37ºC in LB medium containing chloramphenicol (donor plasmid libraries) or ampicillin (recipient plasmid libraries). The complexity of the libraries was estimated by plating the transformed cells on LB agar containing the appropriate antibiotic and found to be $10^6$ to 6 x $10^7$. At least 25 colonies from each library were analyzed by PCR and sequencing and found to have unique combinations of randomized nucleotide residues.

### 3.7.4 Intron mobility and selection experiments

For mobility assays with the full-length EcI5 and Ll.LtrB introns, donor and recipient plasmids were co-transformed into *E. coli* HMS174(DE3), and cells were grown overnight at 37ºC in LB medium containing ampicillin and chloramphenicol. The cells were then diluted 1/100 into fresh LB medium with the same antibiotics, grown at 37ºC to early log phase ($O.D._{595}$ = 0.2-0.3), diluted ten-fold into fresh LB medium, and induced with IPTG under conditions specified in the figure legends for individual experiments. To detect intron integration by PCR, plasmid DNA was isolated from 5 ml of the induced culture by using a QIAprep Spin Miniprep Kit (Qiagen), and the 5'-integration junction was amplified by PCR with the primer RECSEQ01R and an intron-specific primer, either EcI5MOB01 for EcI5 or LtrBAs2.2 for Ll.LtrB (Table 3.2). For quantitative mobility assays using the $Tet^R$ selection system, serial dilutions of the cells were plated onto LB

agar containing tetracycline + ampicillin, or ampicillin alone. Mobility efficiencies were determined as the ratio of $Tet^R$ + $Amp^R$/$Amp^R$ colonies (Guo et al. 2000).

For selection experiments, donor and recipient plasmids or libraries were electroporated into HMS174(DE3), and the transformants were allowed to recover by growth in 1-ml SOC media for 1 h at 37ºC. A small portion of the culture (10 μl) was then removed, serially diluted, and plated on LB agar containing chloramphenicol and ampicillin to yield unselected clones that were sequenced via colony PCR to determine nucleotide frequency biases in the libraries. The remainder of the culture was diluted with 9-ml of LB medium containing ampicillin, chloramphenicol, and 20 mM glucose, the latter being added to decrease leaky expression of the intron from the T7lac promoter. After growing overnight at 37ºC, the cells were harvested by centrifugation, washed with LB medium without glucose, and resuspended in 10 ml of fresh LB medium without glucose. In the experiment of Figure 3.3, a 5-ml portion of the culture was inoculated into 20 ml of LB medium containing 100 μM IPTG and induced for 1 h at 37ºC, while in the experiments of Figures 3.4 to 3.6, a 7-ml portion of the culture was inoculated into 30 ml of LB medium and incubated for 1 h at 37ºC. After the incubations, the cultures were placed on ice for 5 min, and the cells were then harvested by centrifugation, washed twice by centrifugation in ice-cold LB medium if necessary to remove IPTG, and resuspended in 1.5-ml ice-cold LB medium. Finally, the cells were serially diluted, plated on LB agar containing ampicillin, chloramphenicol, and tetracycline, and incubated overnight at 37ºC.

Plasmids containing integrated introns from the selected colonies were either isolated using a QIAprep Spin Miniprep kit and sequenced (Figure 3.3) or analyzed by colony PCR and sequencing (Figures 3.4 to 3.6). Colony PCR was done as described

(Costa and Weiner 2003) with Taq polymerase (0.05 units/μl; New England Biolabs) and 0.4 μM primers in reaction medium containing 20 mM Tris-HCl, pH 8.4, 50 mM KCl, 2.5 mM $MgCl_2$, and 200 μM dNTPs. Reactions were incubated at 94°C for 5 min, followed by 30 cycles of 94°C for 30 sec, 55°C for 30 sec, 72°C for 1 min, and a final extension for 10 min at 72°C.

The nucleotide frequency at each of the randomized positions in the selected introns and target sites was corrected for biases in the initial pool by calculating the ratio $R_{n_p}$ of the frequency of each of the four nucleotides *n* at that position *p* in the selected introns ( $f_{n_p(selected)}$ ) to its frequency in the initial pool ( $f_{n_p(pool)}$ ) by using the equation:

$$\frac{f_{n_p(selected)}}{f_{n_p(pool)}} = R_{n_p}$$

The ratio at each position was then normalized to 1 by using the equation:

$$\frac{R_{n_p}}{\sum_n R_{n_p}} = RN_{n_p}$$

where $RN_{n_p}$ is the normalized ratio. For a given nucleotide residue, a normalized ratio of 0.25 (or 25%) indicates no preference at that position, while larger values indicate selection for the nucleotide residue, and smaller values indicate selection against the nucleotide residue. The normalized frequencies were used to generate a sample set of 100 active target sites and displayed in "WebLogo" format (Crooks et al. 2004).

### 3.7.5 Retargeting EcI5 to insert into different sites

Retargeted EcI5-ΔORF introns were constructed in and expressed from donor plasmids pACD3-EcI5A, C, G, or T, with the choice of donor plasmid dictated by the EBS3 nucleotide residue complementary to the IBS3 nucleotide residue in the DNA target site. The intron was retargeted by a two-step PCR, using the wild-type donor plasmid as template with three unique primers (P1, P2, and P3) and one fixed primer

(P4). P1 corresponds to 5'-exon positions -30 to -14, with modifications at positions -13 to-1 to make IBS1 and IBS2 in the donor plasmid complementary to the retargeted EBS1 and EBS2 for efficient RNA splicing; P2 is complementary to intron positions +238 to +296 with modifications at position +253 to +257 to make EBS2 complementary to IBS2 in the DNA target site; P3 corresponds to intron positions +277 to +336, with modifications at positions +316 to +320 (5 bp) or +321 (6 bp) to make the EBS1 complementary to IBS1 in the DNA target site; P4 is a fixed primer complementary to intron positions +473 to +505. First, overlapping segments of the donor plasmid were amplified by two parallel PCRs, one using primers P1 and P2, and the other using primers P3 and P4. Then, the gel-purified PCR products from the first PCR step were mixed and amplified with the outside primers P1 and P4 to generate a 549-bp PCR product containing sequences corresponding to the 5' exon and 5' end of the intron (nucleotide positions E1 -30 to intron +505). The final PCR product was purified in a 0.8% agarose gel, digested with XbaI and AvaII, and swapped for the corresponding fragment of the donor plasmid.

For *lacZ* gene targeting, the donor plasmid containing the retargeted intron was electroporated into *E. coli* HMS174(DE3). After recovery by growth in 1-ml of SOC media for 1 h at 37ºC, the cells were diluted with 4 ml of LB medium containing chloramphenicol (25 μg/ml) for pACD3-EcI5 or tetracycline (15 μg/ml) for pBL1-EcI5 and grown overnight at 37ºC. A 50-$\mu$l portion of the overnight culture was then inoculated into 5 ml of fresh LB medium containing the same antibiotic and grown to early log phase ($O.D._{600}$ = 0.2-0.3). For induction, 200 μl of the early log-phase culture was inoculated into 5 ml of LB medium containing 100 $\mu$M IPTG (for pACD3-based donor plasmids) or 4 mM m-toluic acid (for pBL1-based donor plasmids) and incubated

at 37ºC for 3 h. The cells were then washed with fresh LB medium and plated on LB agar containing X-gal (40 mg/l). The *lacZ* targeting frequency was determined by counting blue and white colonies.

### 3.7.6 Southern hybridization

Southern hybridizations were as described (Perutka et al. 2004), using DNA isolated from colonies grown up in LB medium with a Qiagen Genomic DNA Isolation Kit (Qiagen). The blots were hybridized with a $^{32}$P-labeled probe corresponding to intron positions 362-739. The probe was generated by PCR of pACD2-EcI5 using primers Ec501s and Ec5T7-AS (Table 3.2), followed by labeling with a High Prime DNA labeling kit (Roche).

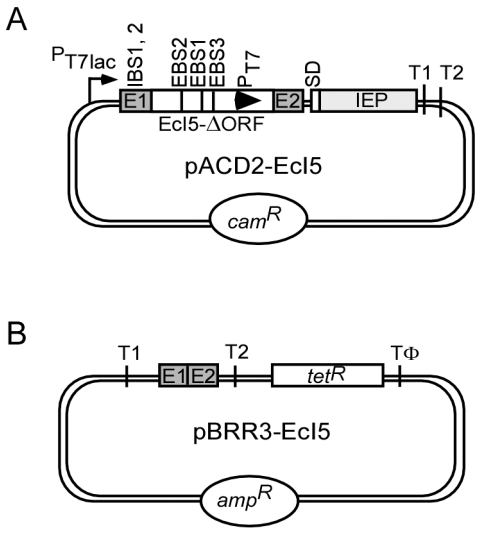



Figure 3.1: EcI5 intron donor and recipient plasmids used in intron mobility assays.

(A) pACD2-EcI5 expresses the EcI5-ΔORF intron and flanking exons, with a phage T7 promoter ($P_{T7}$; arrowhead in intron) inserted in DIV, and the IEP expressed from a position downstream of the 3' exon. SD is the Shine-Dalgarno sequence for the intron ORF. (B) pBRR3-EcI5 contains an EcI5 target site (ligated E1 and E2 sequences) cloned upstream of a promoterless *tet*$^R$ gene. Donor plasmids are derivatives of pACYC184 and carry a *cam*$^R$ marker, while the recipient plasmid is a derivative of pBR322 and carries an *amp*$^R$ marker. T1 and T2 are *E. coli rrnB* transcription terminators. T1 terminates both *E. coli* and phage T7 RNA polymerase, while T2 terminates *E. coli* but not phage T7 RNA polymerase. Tϕ is a phage T7 transcription terminator, which terminates T7 RNA polymerase.

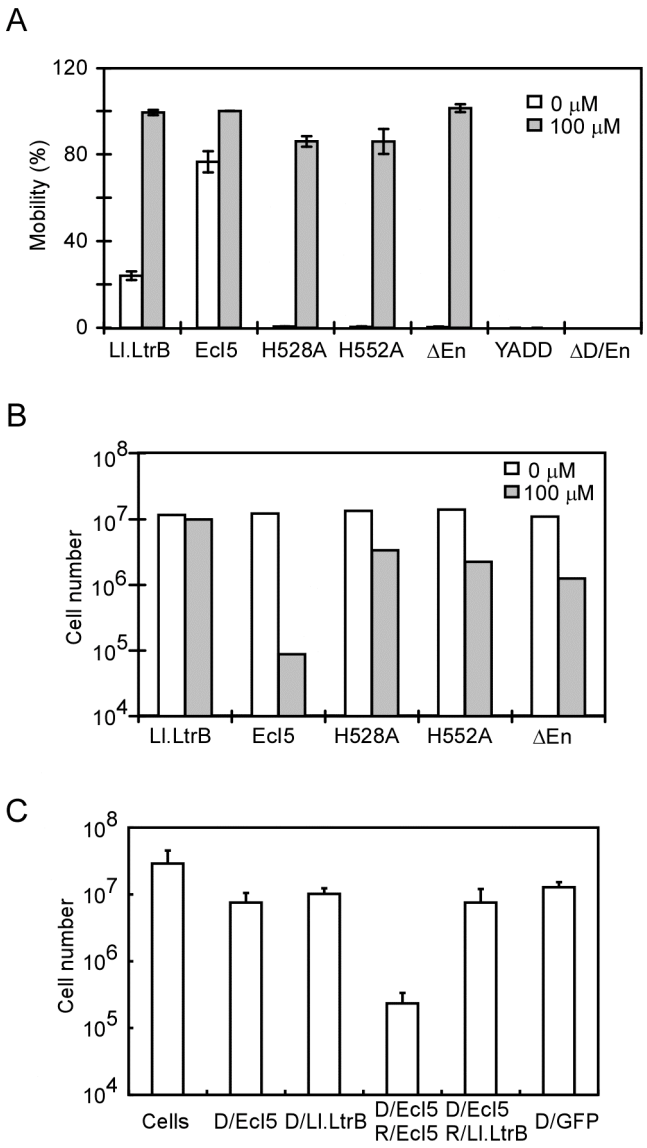



Figure 3.2: Mobility assays for wild-type and mutant EcI5 introns.

(A) Donor plasmids (pACD2X for Ll.LtrB and pACD2-EcI5 or derivatives expressing mutant IEPs for EcI5) and recipient plasmids (pBRR3-ltrB for Ll.LtrB and pBRR3-EcI5 for EcI5) were transformed into *E. coli* HMS174(DE3). The cells were grown at 37ºC to early log phase and incubated with 0 or 100 μM IPTG for 1 h. The introns carry a phage T7 promoter in DIV and integrate into a target site cloned in the recipient plasmid upstream of a promoterless $tet^R$ gene, thereby activating that gene (see Figure 3.1 and Methods). After induction, cells were plated on LB agar containing

ampicillin or ampicillin plus tetracycline, and mobility efficiencies were calculated as the ratio of ($Tet^R$ + $Amp^R$)/$Amp^R$ colonies. The bar graphs show the mean for three determinations, with the error bar indicating the standard deviation. (B) Cell viability after over-expression of the EcI5-ΔORF intron. *E. coli* HMS174(DE3) cells from the mobility assays of panel A were serially diluted after induction and plated on LB medium without antibiotics. (C) *E. coli* HMS174(DE3) with or without the indicated donor and recipient plasmids was grown to early log phase at 37° in LB medium containing the appropriate antibiotics, then induced with 100 μM IPTG for 1 h at 37°C. After induction, the cells were serially diluted and plated on LB medium without antibiotics. The bar graphs show the mean for at least three determinations, except for D/GFP (pAC-GFP; two determinations). The error bars indicate the standard deviation. Abbreviations: D/EcI5, donor plasmid pACD2-EcI5; D/Ll.LtrB, donor plasmid pACD2X; R/EcI5, recipient plasmid pBRR3-EcI5; R/Ll.LtrB, recipient plasmid pBRR3-ltrB; D/GFP, pAC-GFP, a derivative of pACDX that expresses GFP instead of the Ll.LtrB-ΔORF intron and IEP (Zhao and Lambowitz 2005).

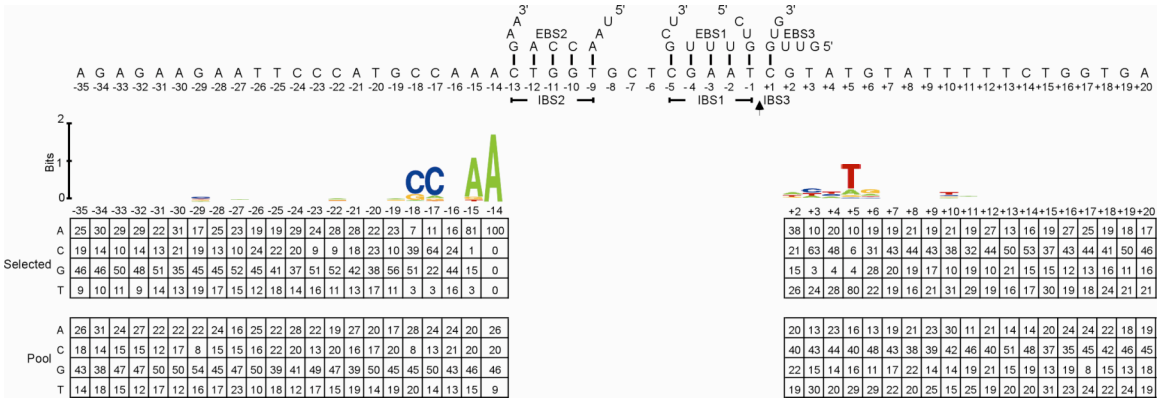



Figure 3.3: Identification of critical nucleotide residues in the distal 5'- and 3'-exon regions of the EcI5 DNA target site.

*E. coli* HMS174(DE3) containing the wild-type intron-donor plasmid pACD2-EcI5 and a recipient plasmid library with random nucleotide residues at positions -35 to -14 and +2 to +20 from the intron-insertion site was grown at 37ºC to early-log phase and induced with 100 μM IPTG for 1 h. The cells were then plated on LB agar containing tetracycline and ampicillin to select those in which the EcI5-ΔORF intron carrying a phage T7 promoter integrated into functional target sequences upstream of the promoterless *tet*$^R$ gene in the recipient plasmid. The targeted plasmids were isolated and sequenced using primer ForpBRR for the 5' integration junction and Rev2pBRR for the 3'-integration junction. The top shows the EcI5 DNA target sequence from positions -35 to +20 and its predicted EBS2/IBS2, EBS1/IBS1, and EBS3/IBS3 base-pairing interactions. The arrow indicates the intron-insertion site. The WebLogo representation (Crooks et al. 2004) depicts nucleotide frequencies at each randomized position in 101 selected target sites, corrected for biases in the initial pool based on sequences of 108 unselected recipient plasmids, as described in Methods. Nucleotide frequencies (percent) at each position in the selected and unselected plasmids are summarized below. In some cases, percentage totals do not equal 100 due to rounding off. This experiment was done by Michael Karberg.

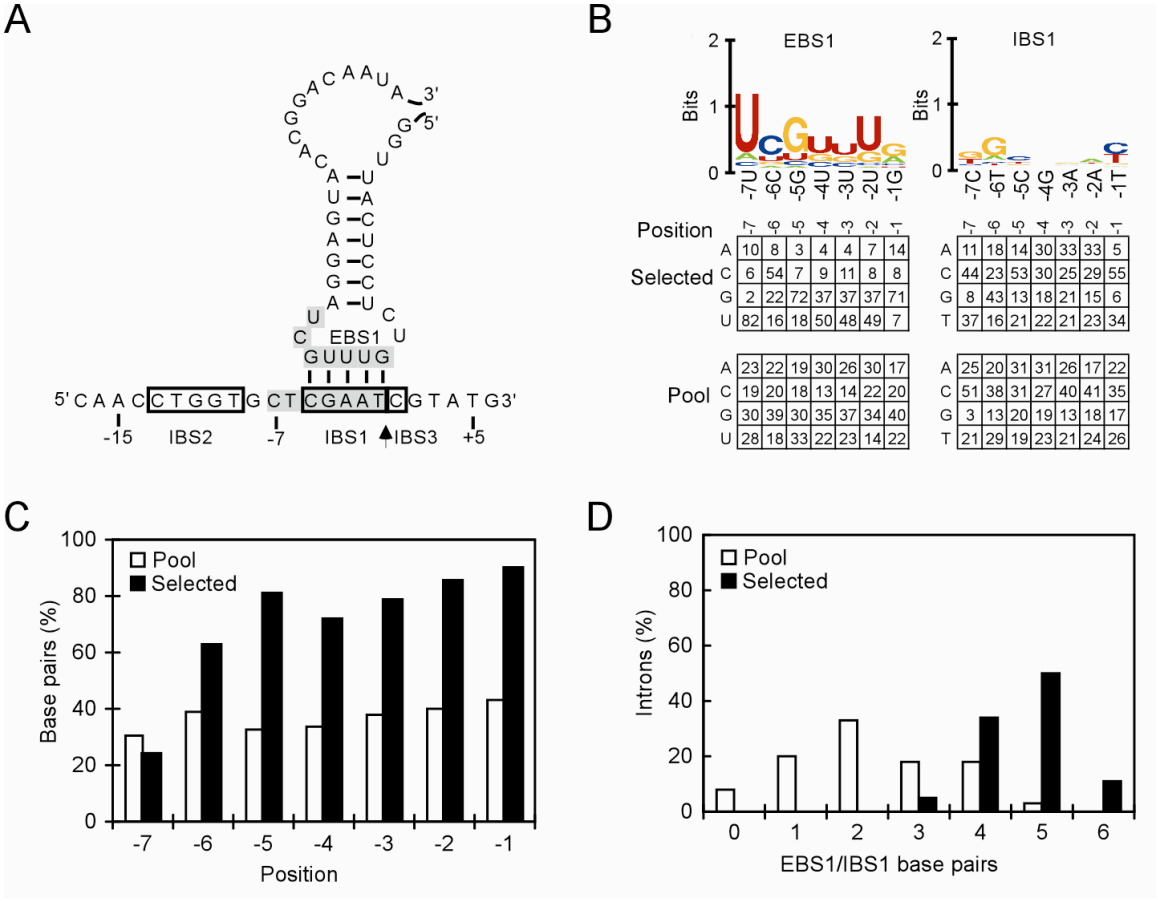



Figure 3.4: Analysis of EcI5 DNA target site EBS1/IBS1 base-pairing interactions by *in vivo* selection.

A selection experiment was done using a pBRR3-EcI5-based recipient-plasmid library with random nucleotide residues at DNA target site positions -7 to -1, and a pACD2-EcI5 based donor-plasmid library with random nucleotide residues at the corresponding positions in the EBS1 stem-loop. The IBS1 positions in the 5'-exon of the donor plasmid were also randomized to provide complementary nucleotide combinations for RNA splicing. The selection was done as in Figure 3.3, except that IPTG induction of the donor plasmid library, which decreased cell viability, was replaced with a 1-h incubation at 37°C in the absence of IPTG. After plating the cells on LB agar containing tetracycline, ampicillin, and chloramphenicol, colonies containing recipient plasmids with integrated introns were picked, and the region from 5'-exon position -210 to intron position +505 was amplified by colony PCR using primers ForpBRR and EcI5AVA2AS and sequenced using primer Rseq (see Table 3.2). (*A*) Predicted EBS1/IBS1 base-pairing interactions between the intron RNA and the top strand of the DNA target site. The nucleotide residues randomized in the selection are highlighted in gray. The arrow

indicates the intron-insertion site. (*B*) WebLogo representation of nucleotide frequencies at randomized positions in 87 selected integration events, corrected for biases in the initial pools based on sequences of 94 unselected donor and 95 recipient plasmids from the initial pools. Nucleotide frequencies (percent) at each position in the selected and unselected plasmids are summarized below. In some cases, percentage totals do not equal 100 due to rounding off. (*C*) Percentage of base pairs at each EBS1/IBS1 position in 87 selected integration events (black bars) and 94 randomly paired donor and recipient plasmids from the initial pools (white bars). (*D*) Percentage of introns that have different numbers of EBS/IBS1 base pairs with the target site in 87 selected integration events (black bars) and 94 randomly paired donor and recipient plasmids from the initial pools (white bars). In (C) and (D), both Watson-Crick and wobble U•G and G•T pairs are counted as base pairs (Guo et al. 1997; Sugimoto et al. 2000).

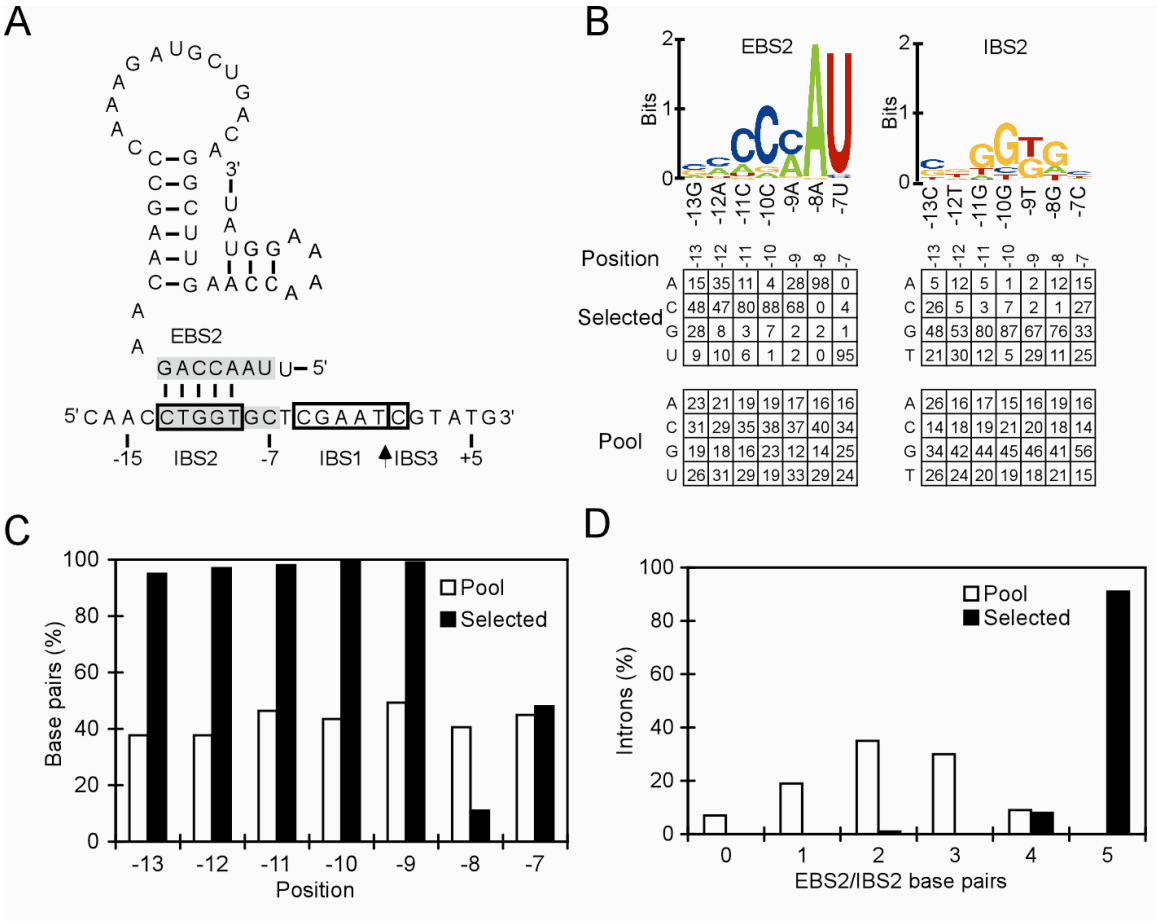



Figure 3.5: Analysis of EcI5 DNA target site EBS2/IBS2 base-pairing interactions by *in vivo* selection.

A selection experiment was done as in Figure 3.4, using a pBRR3-EcI5-based recipient plasmid library with random nucleotide residues at DNA target site positions -13 to -7, and a pACD2-based donor plasmid library with random nucleotide residues at the corresponding positions in the EBS2 loop. The IBS2 positions in the 5'-exon of the donor plasmid were also randomized to provide complementary nucleotide combinations for RNA splicing. Colonies were selected and plasmid sequences determined via colony PCR, as described in Figure 3.4. (*A*) Predicted EBS2/IBS2 base-pairing interactions between the intron RNA and the top strand of the DNA target site. The nucleotide residues randomized in the selection are highlighted in gray. The arrow indicates the intron-insertion site. (*B*) WebLogo representation of nucleotide frequencies at randomized positions in 102 selected integration events, corrected for biases in the initial pools based on sequences of 99 unselected donor plasmids and 103 unselected recipient plasmids. Nucleotide frequencies (percent) at each position in the selected and unselected

plasmids are summarized below. In some cases, percentage totals do not equal 100 due to rounding off. (*C*) Percentage of base pairs at each EBS2/IBS2 position in 102 selected integration events (black bars) and 99 randomly paired donor and recipient plasmids from the initial pools (white bars). (*D*) Percentage of introns that have different numbers of EBS2/IBS2 base pairs in 102 selected integration events (black bars) and 99 randomly paired donor and recipient plasmids from the initial pools (white bars). In (C) and (D), both Watson-Crick and wobble U•G and G•T pairs are counted as base pairs.

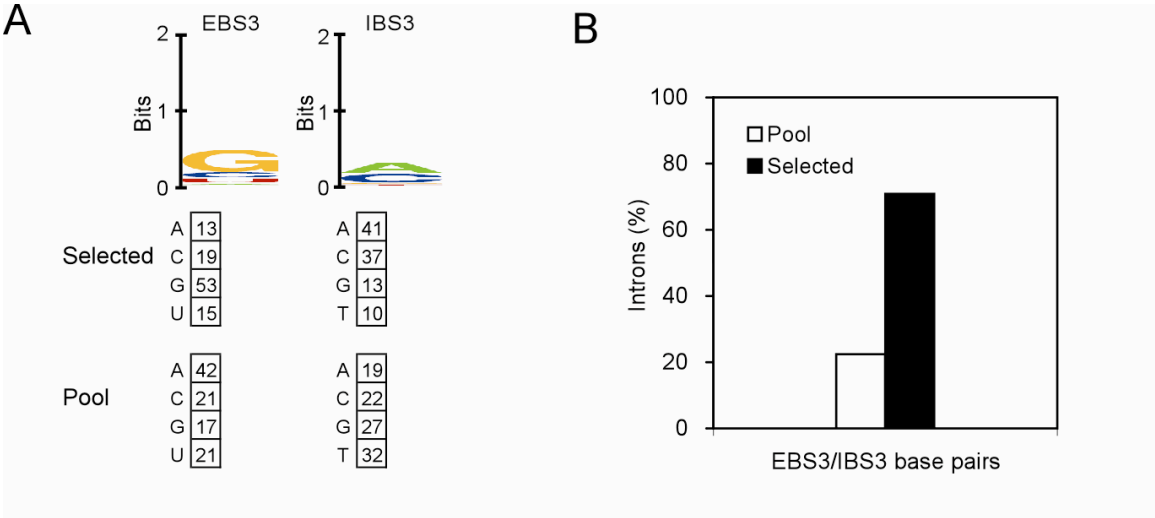



Figure 3.6: Analysis of EcI5 DNA target site EBS3/IBS3 base-pairing interactions by *in vivo* selection.

A selection experiment was done as in Figure 3.4 with pBRR3-EcI5 randomized at IBS3 position +1 and pACD2-EcI5 randomized at EBS3 position +1. After selecting colonies containing recipient plasmids with integrated introns, the region extending from 75-bp upstream of EBS3 to 500-bp downstream of the inserted intron was amplified by colony PCR using primers EcI5 279s and Rev2-pBRR and sequenced using primer EcI5 297s (see Table 3.2). (*A*) WebLogo representation of nucleotide frequencies in 79 selected integration events, corrected for biases in the initial pools based on sequences of 59 unselected donor plasmids, and 59 unselected recipient plasmids. Nucleotide frequencies (percent) at the randomized position are summarized below. In some cases, percentage totals do not equal 100 due to rounding off. (*B*) Percentage of introns having Watson-Crick or wobble U•G and G•T base pairs at EBS3/IBS3 in the 79 selected integration events and 59 randomly paired donor and recipient plasmids from the initial pools.

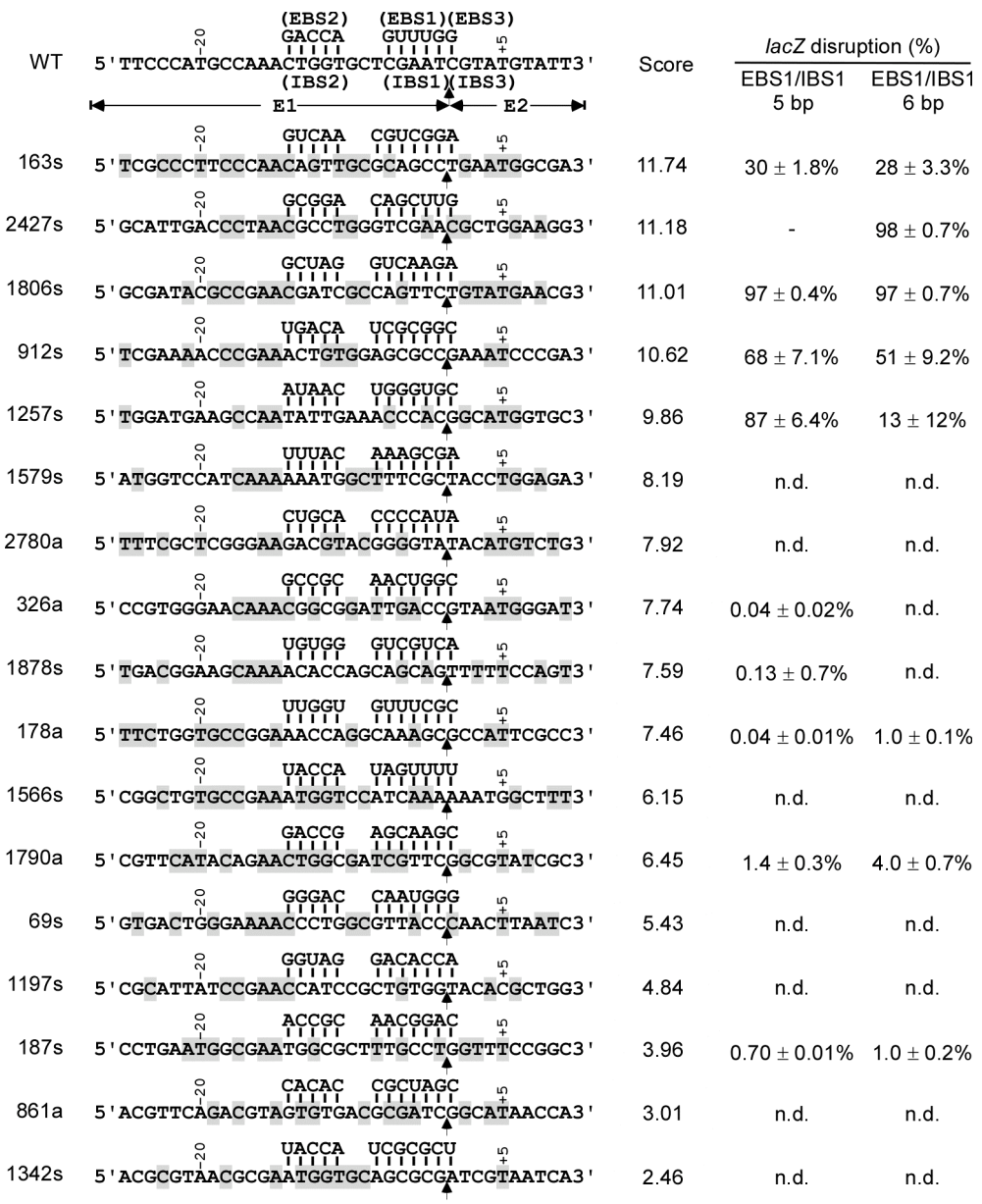



Figure 3.7: DNA target site sequences and base-pairing interactions for EcI5 introns retargeted to insert at different sites in the *E. coli lacZ* gene.

The retargeted EcI5 introns (targetrons) are denoted by a number that corresponds to the nucleotide position 5' to the intron-insertion site in the target gene's coding sequence, followed by "a" or "s" indicating antisense or sense-strand, respectively. DNA target sequences in the *lacZ* gene are shown from positions -26 to +10 from the intron-insertion site, with nucleotide residues that match those in the wild-type (WT) EcI5 target site highlighted in gray. The arrow indicates the intron-insertion site. Log-odds scores

determined using the computer algorithm described in the text are indicated to the right, along with the targeting frequency determined by plating on LB agar containing X-gal and counting blue and white colonies. Targeting frequencies were determined for targetrons having 5 or 6 bp EBS1/IBS1 interactions. The values are the mean ± the standard deviation for three determinations. "-", not determined; n.d., not detectable.

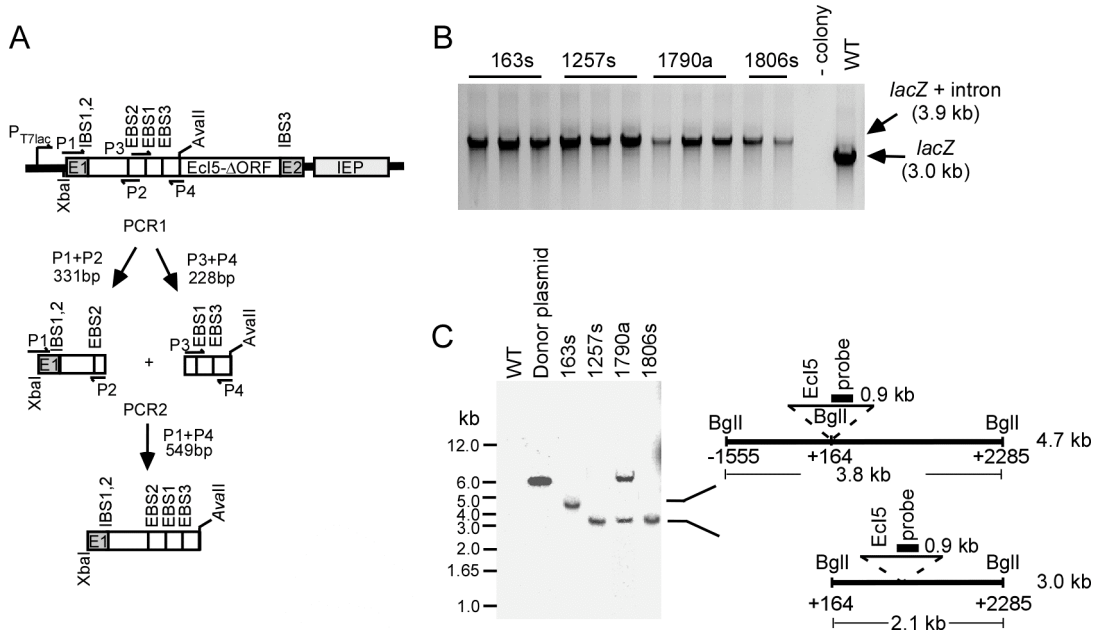



Figure 3.8: Retargeting of EcI5 to insert into different sites in the *E. coli lacZ* gene.

*E. coli* HMS174(DE3) was transformed with pACD3-EcI5 expressing the retargeted introns, grown at 37ºC to early-log phase, and induced with 100 μM IPTG for 3 h unless noted otherwise. The cells were then plated on LB agar containing X-gal (40 mg/l), and the *lacZ* targeting frequency (Figure 3.7) was determined by counting blue and white colonies. (*A*) Two-step PCR used to retarget the EcI5-ΔORF intron by modification of EBS and IBS sequences in the donor plasmid. The donor plasmid used as template was pACD3-EcI5A, C, G, or T according to the desired EBS3 nucleotide residue. P1-P4 are primers used to modify the intron's EBS1 and 2 to be complementary to IBS1 and 2 in the DNA target site, and IBS1 and 2 in the 5' exon of the donor plasmid to be complementary to the retargeted EBS1 and EBS2 for efficient RNA splicing (see Methods). The final PCR product containing the modified sequences was digested with XbaI and AvaII and swapped for the corresponding fragment of the same donor plasmid. (*B*) Colony PCR of *E. coli lacZ* disruptants obtained using retargeted EcI5 introns. Colony PCR was done with primers LacZP3 and LacZP4 flanking the intron-insertion site in the *lacZ* gene (see Table 3.2). The figure shows representative data for three of the retargeted introns. "- colony" is a parallel PCR without a colony, and WT is a parallel PCR done on a colony of wild-type HMS174(DE3). (*C*) Southern hybridizations. Genomic DNA was isolated from *E. coli* HMS174(DE3) (WT) and the indicated l*acZ* disruptants grown under non-selective conditions. The DNA was digested with BglI, run in a 0.8% agarose gel, blotted to a nylon membrane, and hybridized with a $^{32}$P-labeled intron probe (see Methods). The donor plasmid pACD2-EcI5 digested with BglI was run in a parallel lane. The numbers to the left of the gel indicate the positions of size markers

(1-kb plus ladder; Invitrogen). The schematics to the right depict the BglI fragments of the *lacZ* gene containing the inserted EcI5 intron. The insertion of targetron LacZ1257s, 1790a, or 1806s results in a 3.0-kb BglI fragment. The insertion of targetron LacZ163s disrupts the BglI sites at position 164, resulting in a larger fragment (4.7 kb). The LacZ1709a disruptant shows an additional band due to residual donor plasmid, which was retained in this disruptant but lost in the other disruptants during growth under non-selective conditions.

| | 3' | | | | | 5' | | 3' | | | | | 5' | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ( | EBS2 | | | | ) | | ( | EBS1 | | | | ) | EBS3 |
| EcI5 RNA | G | A | C | C | A | A | C | G | U | U | U | G | | G |
| DNA target site | C | T | G | G | T | G | T | C | G | A | A | T | | C |
| | IBS2 | | | | | | | IBS1 | | | | | | IBS3 |
| RNA/DNA | -13 | -12 | -11 | -10 | -9 | -8 | -6 | -5 | -4 | -3 | -2 | -1 | | +1 |
| A/A | 0 | 0 | 0 | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | | 4 |
| A/C | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | | 0 |
| A/G | 0 | 0 | 0 | 0 | 0 | 74 | 1 | 0 | 0 | 0 | 1 | 0 | | 0 |
| A-T | 15 | 28 | 11 | 4 | 28 | 11 | 6 | 2 | 3 | 3 | 3 | 14 | | 9 |
| C/A | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | | 9 |
| C/C | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 1 | 0 | | 2 |
| C-G | 47 | 51 | 79 | 86 | 66 | 0 | 34 | 5 | 8 | 10 | 7 | 6 | | 9 |
| C/T | 1 | 2 | 1 | 1 | 1 | 0 | 9 | 0 | 1 | 1 | 0 | 1 | | 0 |
| G/A | 1 | 0 | 1 | 0 | 0 | 0 | 3 | 6 | 6 | 6 | 1 | 2 | | 13 |
| G-C | 26 | 5 | 3 | 7 | 2 | 0 | 13 | 49 | 23 | 23 | 23 | 49 | | 36 |
| G/G | 0 | 0 | 0 | 1 | 0 | 2 | 5 | 6 | 2 | 2 | 0 | 0 | | 0 |
| G•T | 2 | 1 | 0 | 0 | 0 | 0 | 2 | 14 | 6 | 7 | 14 | 20 | | 2 |
| U-A | 5 | 12 | 5 | 1 | 2 | 0 | 7 | 8 | 24 | 27 | 32 | 2 | | 16 |
| U/C | 0 | 0 | 0 | 0 | 1 | 0 | 7 | 2 | 5 | 2 | 3 | 5 | | 0 |
| U•G | 1 | 1 | 1 | 0 | 0 | 1 | 2 | 2 | 8 | 8 | 7 | 0 | | 0 |
| U/T | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 13 | 9 | 7 | 0 | | 0 |
| Base paired (%) | 95 | 98 | 98 | 98 | 98 | 12 | 64 | 80 | 72 | 78 | 85 | 91 | | 72 |

Table 3.1: Observed base-pair frequencies in selection experiments for the EBS1/IBS1, EBS2/IBS2, and EBS3/IBS3 interactions between the EcI5 intron RNA and its DNA target site.

The table shows the percentage of introns having the indicated RNA-DNA base pair or mismatches at each position in the EBS1/IBS1, EBS2/IBS2, and EBS3/IBS3 DNA target site interactions in the selection experiments of Figures 3.4 to 3.6. The wild-type base pair at each position is boxed. Base-pairing interactions between wild-type EcI5 and its DNA target site are shown at the top.

| Oligonucleotide | Sequence (5' to 3') |
|---|---|
| DelD/EnA | GAACACCGACTAGTCATTGATGGAAATATCTTTTCCGAATCCAC |
| DelEnA | GAACACCGACTAGTCATCGGCCCCGCTGCTCAAAC |
| DelEnS | GACAATGAGATCTGGCTGGCTTTATGG |
| E2-3'ORF | CTGCAGCCCCGCCCCCATTTAAATATATCTCGAGCACCCGTTCTCGGAGC |
| EBS3N | CCACCGGTGTTNTGAGCAGAGTCTAACC |
| Ecl5 279s | GGCTTGAACCAAAAGGTATG |
| Ecl5 3' | GTACCCAACCTCGAGAACACCGACTAGTCAAGCCT |
| Ecl5 347a | GGTGGTTTATTGTCCGTGTACCCCTAGCAAACA |
| Ecl5 5' | CCCCAAGCTTAGAATTCCCATGCCAA |
| Ecl5 IBS1A | GCGGCGGAATTCTCACCAGAAAAATACATACGNNNNNNNCACCAGTTTGGCATG |
| Ecl5 IBS1S | GCGGCGGACGTCAGAGGAGAACTCCCATGCCAAACTGGT |
| Ecl5 IBS2A | GCGGCGGAATTCTCACCAGAAAAATACATACGATTCG |
| Ecl5 IBS2S | GCGGCGGACGTCAGAGGAGAACTCCCATGCCAAANNNNNNNNTCGAATCGTATGTAT |
| Ecl5 IBS3A | GCGGCGGAATTCTCACCAGAAAAATACATACNATTCGAGCACCAGTTTGGCATG |
| Ecl5 IBS3S | GCGGCGGACGTCAGAGGAGAACTCCCATGCCAAACTGGT |
| Ecl5-01S | AGTCTAACCTACTGTTGTTATTCAGGTGGAA |
| Ecl5-3'exon | GGGTTTAAACGGGGCTGCAGAGAAAAATACATACG |
| Ecl5-3'N19 | CCCGAATTCNNNNNNNNNNNNNNNNNNNGATTCGAGCACCAG |
| Ecl5-3'ORF | CCAACCTCGAGAACACCGACTAGTCAAGCCTTTCTAAGCCCACTCTCATG |
| Ecl5-5'exon | CCCCTCTAGAAGAATTCCCATGCCAAACTGGTGCTCGAAT |
| Ecl5-5'N22 | AAAGACGTCCGGCGGNNNNNNNNN NNNNNNNNNNNNNNCTGGTGCTCGAATC |
| Ecl5-5'ORF | CCCCGTTTAAACCCTAAGAAGGAGATATACCTATGACTGAGCAGGCTACAACCTGTAAA |
| Ecl5AVA2AS | ATCCGGTCCATTACAGACTGGCATTCGCTTTTT |
| Ecl5BASWT | ATACCTTTTGGTTCAAGCCTGTCAGCATCTTTG |
| Ecl5EBS1-N | GGCTTGAACCAAAAGGTATGTGGTTGGTTACTCCTCTNNNNNNNAGGGGTACACGGACAATA |
| Ecl5EBS1WT | GGCTTGAACCAAAAGGTATGTGGTTGGTTACTCCTCTGTTTGCTAGGGGTACACGGACAATA |
| Ecl5EBS2-N | ATACCTTTTGGTTCAAGCCTGTCAGCATCTTTGGCTTGTTNNNNNNNACGACGCTTC |
| Ecl5EBS3N | CTAGGGGTACACGGACAATAAACCACCGGTGTTNTGAGCAGAGTC |
| Ecl5IBS1-N | CCCCTCTAGAAGAATTCCCATGCCAAACTGGTGNNNNNNNGTGCGACATGAAG |
| Ecl5IBS2-N | CCCCTCTAGAAGAATTCCCATGCCAAANNNNNNNTCGAATGTGCGACATGAAG |
| Ecl5MOB01 | CTTTTGGTTCAAGCCTGTCAGCATC |
| Ecl5-P1 | ACGCGTTTGCCTTTACGATACGCGCTT |
| Ecl5-P3 | GTATCGTAAAGGCAAACGCGTCACTTGTTCATGCGAAA |
| Ecl5T7-AS | ACGCGTTTGCCTTTACGATACGCGCTT |
| ForpBRR | CTGATCGATAGCTGAAACGC |
| IBS3N | GCATCACTGCTACCCGATNGTATGTATTTTTCTC |
| LacZP3 | CGGATAACAATTTCACA |
| LacZP4 | ACATGGCCTGCCCGG |
| LtrBAs2.2 | TTCTCCTACAGATTGTACAAATGTGG |
| pACD229s | GCGTCCGGCGTAGAGGATCG |
| RECSEQ01R | GGGGTTCCGCGCACATTTCCCCGAAAAGTGCC |
| Rev2pBRR | AATGGACGATATCCCGCA |
| Rseq | CCATGCGAGAGTAGGGAAC |
| T7MLU1-B | CGCGTCCCTATAGTGAGTCGTATTATA |
| T7MLU1-T | CGCGTATAATACGACTCACTATAGGGA |

Table 3.2: DNA oligonucleotides used in Chapter 3.

DNA oligonucleotides were synthesized by Integrated DNA Technologies.

# Chapter 4: Gene targeting in *Xenopus laevis* via site-specific modification of sperm DNA using mobile group II introns

Mobile group II introns found in bacterial and organelle genomes have been developed into highly efficient bacterial gene targeting vectors ("targetrons") (Lambowitz and Zimmerly 2004). The utility of mobile group II introns for gene targeting derives from their unique site-specific DNA integration mechanism in which the intron RNA reverse splices directly into a DNA strand and is then reverse transcribed by its intron-encoded protein (IEP). This process is mediated by a ribonucleoprotein particle (RNP) that contains the IEP and the excised intron RNA. Both the IEP and intron RNA are involved in DNA target site recognition, with most of the specificity coming from base pairing of the intron RNA to the target sequence (Eskes et al. 1997; Guo et al. 1997; Guo et al. 2000; Singh and Lambowitz 2001). Thus, it is possible to retarget the intron to insert into desired DNA sites simply by modifying the base pairing sequences in the intron RNA (Guo et al. 2000; Mohr et al. 2000; Perutka et al. 2004; Lambowitz et al. 2005).

The Ll.LtrB group II intron has been used as a model system to study group II intron mobility and to develop gene targeting applications (Lambowitz and Zimmerly 2004). The catalytically active intron RNA consists of six conserved domains (DI-VI; Figure 4.1A; (Michel and Ferat 1995; Lehmann and Schmidt 2003). The IEP (denoted LtrA protein), which is encoded in a large "loop" in DIV, binds specifically to the intron RNA to stabilize the catalytically active RNA structure and form RNP particles. The RNPs can carry out site-specific DNA integration by reverse splicing. RNPs initiate DNA integration by binding to DNA target sites, with the LtrA recognizing a small number of bases in the distal 5'-exon region and triggering local DNA melting, enabling intron RNA

sequences denoted EBS2, EBS1, and δ to base pair to sequences denoted IBS2, IBS1, and δ' spanning a 13-nt region of the DNA target sites (Figure 4.1B). After base pairing, the intron RNA reverse splices into one strand of the DNA, while LtrA cleaves the opposite strand a short distance downstream (between position +9 and +10) and uses the cleaved 3' end as a primer for reverse transcription of the inserted intron RNA (Cousineau et al. 1998; Smith et al. 2005). The retargeting of the Ll.LtrB intron into different sites can be achieved simply by modifying the EBS2, EBS1, and δ sequences in the intron RNA. Its high insertion efficiency and specificity have made it possible to develop the Ll.LtrB intron into a highly efficient gene targeting vector (targetron) for use in diverse bacteria (Karberg et al. 2001; Perutka et al. 2004; Yao et al. 2006).

In principle, mobile group II introns can be used for gene targeting in eukaryotes, but challenges include getting a sufficiently high concentration of active group II RNPs into the nucleus and the chromatin structure of DNA target sites (Mastroianni et al. 2008). One way to circumvent these problems is to use group II introns for *in vitro* modification of chromosomal DNA in decondensed sperm nuclei, which can then be used for *in vitro* fertilization to obtain transgenic animals. Procedures for generating transgenic animals via random modification of sperm DNA are well established in *X. laevis* (Kroll and Amaya 1996; Amaya and Kroll 1999; Sparrow et al. 2000), and analogous approaches have been developed for zebrafish (Jesuthasan and Subburaju 2002), mice (Perry et al. 1999; Kaneko et al. 2005; Smith and Spadafora 2005), livestock (Rieth et al. 2000; Kaneko et al. 2005; Smith and Spadafora 2005) and monkeys (Chan et al. 2000b; Chan et al. 2000a).

For *X. laevis*, the standard sperm transgenesis protocol involves mixing permeabilized sperm nuclei with a linear DNA fragment, then partially digesting sperm

genome with a restriction enzyme, and injecting the mixture into eggs (Kroll and Amaya 1996; Amaya and Kroll 1999; Sparrow et al. 2000). The exogenous DNA then inserts into the double-strand DNA breaks introduced by the restriction enzyme and integrates into chromosomal DNA by host cell DNA repair mechanisms. The procedure is very rapid, enabling the generation of hundreds of transgenic embryos per day. With current methods, however, the DNA fragment integrates randomly throughout the genome and multiple copies of DNA can insert at different locations. Consequently, this technique has been limited to gain-of-function experiments. By adding retargeted group II RNPs and eliminating the restriction enzyme from the *Xenopus* transgenesis protocol, it is possible to integrate the intron site-specifically rather than randomly into sperm nuclei (Figure 4.2). By using the targeted sperm nuclei for *in vitro* fertilization, one can then generate site-specific transgenic animals.

In this chapter, I show that the mobile *Lactococcus lactis* Ll.LtrB group II intron (Ll.LtrB) can site-specifically insert at desired chromosomal sites in *Xenopus laevis* sperm nuclei, and that the resulting modified sperm nuclei can then be used for *in vitro* fertilization to generate genetically modified animals. Because the targeted gene modification occurs in a germ line cell prior to fertilization, the resulting embryos and animals are not chimeric. These results constitute proof-of-principle that group II introns can be used for gene targeting in eukaryotes via sperm DNA modification.

## 4.1 SELECTION OF LL.LTRB TARGET SITES IN *XENOPUS*

Ll.LtrB intron target sites in *X. laevis* were chosen by a computer algorithm. First, the algorithm scans a gene sequence for the best matches to positions recognized by the LtrA. It then designs primers for PCR modification of the EBS1, 2, and δ sequences

within the intron RNA to base pair optimally to the DNA target site for efficient intron integration (Perutka et al. 2004).

The required sequence modifications were introduced into an intron expression construct, pACD2 or pACD3, which contain a 0.9-kb Ll.LtrB-ΔORF intron and flanking exons, cloned behind a phage T7 promoter. Another T7 promoter was inserted into intron DIV in pACD2 but not in pACD3. The targeting efficiency of the modified introns was assessed both by an established *E. coli* mobility assay *in vivo* (Guo et al. 2000) and by a target DNA-primed reverse transcription assay (TPRT assay) *in vitro* (Saldanha et al. 1999). For the *in vivo* assay, the modified pACD2 donor plasmid expressing the retargeted intron with a phage T7 promoter in intron DIV was introduced into *E. coli* HMS174(DE3) with an ampicillin-resistant recipient plasmid containing the intron's target site (Guo et al. 2000). The intron target site is located upstream of a promoterless $tet^R$ gene, and the insertion of intron into the target site activates the expression of that gene. Thus the targeting efficiency of modified introns can be determined as the ratio of tetracycline-resistant + ampicillin-resistant colonies to ampicillin-resistant colonies.

For the *in vitro* TPRT assay, the intron RNA precursor is transcribed directly from the linearized pACD3 vector and self-spliced under conditions of high salt and $Mg^{2+}$ concentration (Matsuura et al. 2001). The spliced RNA is then reconstituted with purified LtrA protein to form RNPs that promote DNA integration. During the *in vitro* TPRT reaction, plasmids containing the DNA target site are incubated with reconstituted RNPs, dNTPs and [$^{32}$P]-labeled dTTP. The [$^{32}$P]-labeled dTTP is then incorporated into cDNA copies of the reverse spliced products, which are then visualized by using a phosphorimager.

Several *X. laevis* genes (28S rRNA, transposable element Tx1, tyr and mitf) were chosen. The 28S rRNA gene and Tx1 gene are multiple-copy genes in the *X. laevis*. There are about 200-300 copies of the 28S rRNA gene per genome (Roger et al. 2002) and about 150 copies of the Tx1 transposon per genome (Garrett and Carroll 1986). The existence of multiple copies in the genome provides a higher chance for intron integration. In addition, two single-copy genes, tyr, encoding tyrosinase (Tyr), and mitf, encoding microphthalmia-associated transcription factor (Mitf) were chosen (Kumasaka et al. 2003; Kumasaka et al. 2004). Tyrosinase is required for melanin production in melanocytes, and homozygous mutants are albino in medaka fish (Koga et al. 1995), mice (Rinchik et al. 1993), and humans (Oetting and King 1993). Mitf positively regulates the transcription of pigmentation-related genes (*e.g.*, tyr, trp1, oa1) in several cell lineages, including melanocytes and retinal pigment epithelial cells in mice and zebrafish (Lister et al. 1999; Gaggioli et al. 2003). In *Xenopus*, Mitf is expressed in premigratory melanoblasts and retinal pigment epithelium (Kumasaka et al. 2004). Mice and frogs have two Mitf isoforms, Mitf-A and Mitf-M, which are produced by alternative splicing in the eye and body, respectively. Further, in *X. laevis* (but not *X. tropicalis*), the mitf gene is duplicated, the two genes being referred to as mitf-α and mitf-β. In mice, a dominant negative mitf allele has been shown to down-regulate tyrosinase expression in melanocytes (Vachtenheim and Drdova 2004). In *Xenopus*, expression of a dominant negative mitf-α results in failure of differentiation and pigmentation of the eye (Kumasaka et al. 2005). The tyr and mitf genes were chosen because they have easily visible, localized expression patterns, facilitating detection of intron insertion with GFP marker strategies (see section 4.6) (Bronchain et al. 1999) and because homozygous

mutants are viable with phenotypes that are readily scorable in embryos by loss of pigmentation.

The DNA target site sequences in the four chosen genes are shown in Figure 4.3. For single copy protein coding genes, targetrons were designed to insert specifically into tyr exon 2 (targetron Tyr-401a) and mitf-α-M exon 6 (targetron Mitf-α-M-609s). The targetrons were validated by the *in vivo E. coli* plasmid-based mobility assay. The mobility frequencies were 86%, 59%, 56% and 65% for the targetron 28S-3205s, Tx1-2768a, Tyr-401a and Mitf-α-M-609s, respectively. Additionally, the reverse splicing activity of each RNP was assessed by *in vitro* TPRT assay. Figure 4.3C showed that the Mitf-α-M-609s, Tyr-401a and Tx1-2768a targetron RNPs can site-specifically reverse splice into their own target site and promote target DNA-primed reverse transcription into their own target plasmids (similar results were obtained for the 28S-3205s intron in a separate *in vitro* TPRT assay). Both full and partial reverse splicing products were detected in the reactions (Saldanha et al. 1999; Mastroianni et al. 2008). Smaller size cDNA synthesis products were also detected. No product was detected when the retargeted RNPs were incubated with recipient plasmid with a mock target site (data not shown), demonstrating the specificity of the retargeted introns.

### 4.2 SPERM TARGET DNA-PRIMED REVERSE TRANSCRIPTION (TPRT) ASSAY

The group II intron-mediated gene targeting method developed for *X. laevis* is based on the current transgenesis protocol. As outlined in Figure 4.2, the sperm nuclei were incubated with *in vitro* reconstituted RNP particles, nucleoplasmin, dNTPs and $Mg^{2+}$ in sperm dilution buffer (SDB) at specific temperatures to allow intron insertion into the desired target sites in the sperm genome. The sperm nuclei used in the reaction were permeabilized with lysolecithin and the chromatin structure was decondensed

during the reaction using purified recombinant nucleoplasmin (Banuelos et al. 2003). These treatments enable RNPs to enter the sperm nuclei and access the DNA target sites. The intron integration can be checked right after the sperm reaction. This is important as numerous factors need to be tested to optimize the intron targeting conditions. Although the same thing can be done with oocytes, the collection of oocytes is very laborious compared to sperm. The amount of sperm that can be collected from a male frog is thousands of times more than that of oocytes collected from a female one.

To check the targeting efficiency, the sperm reaction was terminated by heating at 75ºC for 10 min. If intron-mediated integration occurs at the chosen target site, cDNA synthesis initiates in the presence of dNTPs after the intron reverse splicing. The newly synthesized cDNA joined with the bottom strand of intron 3'-exon can be detected by PCR. However, the cDNA and intron 5'-exon junction does not exist in sperm reactions either because of incomplete cDNA synthesis or the lack of DNA repair enzymes in sperm (Derijck et al. 2006).

To detect the intron 3'-integration junction, a sense strand primer annealing near the 3' end of Ll.LtrB and an antisense strand primer annealing to the target gene sequence located downstream of the intron-insertion site were used for PCR. The PCR amplification of the 3'-integration junction was found to be improved by treating the reacted sperm DNA with RNase H, which implies the existence of an intron RNA and cDNA hybrid after the sperm reaction. The removal of the intron RNA strand by RNase H facilitates the PCR amplification. To determine the intron targeting efficiency in the sperm reaction, the same PCR was applied to the serially diluted sperm reaction DNA. The targeting efficiency was also verified by real-time PCR. For the intron targeting in Tx1, mitf and tyr gene, the 3'-integration junction was detected in as few as 50, 500 and

301 sperm nuclei respectively (Figure 4.4 and 4.7, targeting efficiency for tyr gene is not shown here). I will refer to the sperm reaction followed by PCR analysis as the sperm TPRT assay in this chapter.

During the initial experiments, sperm TPRT reactions were focused on getting high intron targeting efficiency. The sperm nuclei were treated at relatively high temperature (33ºC or above) compared to the frog's optimal living temperature (18ºC), long reaction time (30 min), and extensive chromatin decondensation by nucleoplasmin. It is known that the exposure of sperm nuclei to high temperature can reduce their fertility, change sperm morphology, and alter chromatin packing (Banks et al. 2005). In transgenesis experiments done in our laboratory, extensive chromatin decondensation was found to have similar effects. Thus, it is critical to establish sperm reaction conditions that allow efficient intron integration without damaging the sperm nuclei.

### 4.2.1 The targetron integration efficiency is proportional to the amount of RNPs used in the sperm reaction

The targetron Tx1-2768a RNP was found to successfully target the Tx1 transposon in sperm nuclei. As shown in Figure 4.4B, Tx1-2768a intron integration efficiency was checked with 1, 5 and 10 μl of Tx1-2768a RNP (1.8 mg/ml lariat RNA based on $O.D._{260}$). No targeted product was detected in the reaction using 1 μl of Tx1-27678a RNP (lane 1 and 5). The targeted product was only present in the reactions with 5 or 10 μl of RNP (lane 2, 3, 6 and 7). Furthermore the band intensity of PCR product was proportional to the amount of RNPs used the reaction. A negative control with 10 μl of heat-inactivated RNPs did not yield the same targeted product (lane 4 and 8), confirming that the detected products resulted from the activity of the RNPs. In summary, the active RNP is required for the intron integration into the target site in sperm nuclei and the

integration efficiency is proportional to the amount of active RNPs used in the sperm TPRT reaction.

Two nucleoplasmin protein concentrations (4.7 and 23.7 μM) were tested in the Tx1-2768a intron sperm TPRT reaction (Figure 4.4B). Surprisingly, no significant difference was observed between the two concentrations. This result implies that for the Tx1-2768a targetron, the amount of nucleoplasmin required for maximally efficient intron integration might be 4.7 μM or less. It also demonstrated that excess nucleoplasmin did not increase the intron targeting efficiency. More detailed optimization of nucleoplasmin in the sperm TPRT reaction will be discussed in the section 4.2.3.

In Figure 4.4B lane 9, 1 μl of RNPs was incubated with extracted sperm genomic DNA in the absence of nucleoplasmin. Surprisingly, the targeted product was detected with the extracted genomic DNA, but not with sperm nuclei plus nucleoplasmin (lane 1 and 5). This finding confirms that chromatin structure is one of the obstacles for group II intron targeting in eukaryotes. The same conclusion was reached in *X. laevis* oocyte targeting experiments (Mastroianni et al. 2008).

Next, the Tx1-2768a intron targeting efficiency was determined by the same PCR method using the serially diluted sperm TPRT reaction DNA. During the preparation of sperm nuclei, the amount of sperm nuclei per μl was determined by counting Hoechst stained sperm nuclei on a hemacytometer (Smith et al. 2006). The number of sperm nuclei used in sperm TPRT reaction was known (within the range of $4\sim8\times10^5$ per reaction). For the Tx1-2768a RNP, the targeted product can be detected as few as 54 sperm nuclei (Figure 4.4B). No targeted product was detected in the PCR negative controls, distilled water. The products indicated by the arrow corresponding to the

targeted 3'-integration junction were confirmed by sequencing (Figure 4.4C). All other bands were PCR primer dimers or PCR artifacts.

As shown in Figure 4.4D, the Tx1-2768a intron integration efficiency was later confirmed by real-time PCR. Two sets of primers were designed to specifically amplify the 3'-intron integration junction of the Tx1-2768a targetron (Target) and the single copy mitf gene, as a measure of the number of sperm nuclei. The intron targeting efficiency is equal to the ratio of "Target" copy number to the "mitf" gene copy number. Figure 4.4D showed the Ct values of "Target" and "mitf" from two different amount sperm TPRT reaction DNA. The $\Delta Ct = Ct_{Target} - Ct_{mitf}$ represents the fold of "Mitf" copies over "Target" copies in the sperm TPRT reaction. The targeting efficiency was determined by the equation shown in Figure 4.4D. The targeting efficiency based on the real-time PCR is 12.2%, which is higher than that determined by the dilution experiments in Figure 4.4B. The difference could be due to higher sensitivity of real-time PCR than direct PCR.

In summary, the intron integration in sperm requires active RNP, and higher integration efficiency can be obtained by adding more RNPs to the reaction. Recently improved methods have made it possible to obtain RNP concentrations of almost 10 mg/ml based on the RNA absorbance at $O.D._{260}$. It has not yet been tested how very high concentrations of RNPs in the reaction affect the intron targeting efficiency.

### 4.2.2 The targetron integration efficiency is dependent upon temperature in the sperm reaction

The Ll.LtrB intron can reverse splice efficiently at the growth temperature of its natural host, *L. lactis* (30°C) and the well-studied host, *E. coli* (37°C). However the optimal living temperature for the *X. laevis* is between 18°C and 23°C (Sive et al. 2000). Incubating *X. laevis* sperm nuclei at higher temperatures can damage the integrity of sperm DNA and cause failure in fertilization or embryo development (Paul et al. 2008).

Therefore, it is important to optimize the sperm reaction temperature to allow efficient intron targeting without impairing *in vitro* fertilization and the embryo development.

First, sperm TPRT of the Tx1-2768a targetron was assayed at different temperatures. Earlier transgenesis experiments showed that incubating sperm nuclei above 33ºC dramatically decreased the fertilization efficiency (data not shown). Therefore, the temperature optimization experiments were focused on temperatures ranging from 25ºC to 33ºC. As shown in Figure 4.5, the Tx1-2768a intron targeted product can be detected at all of the above temperatures, but not in the DNA from sperm nuclei that had not been treated with the targetron or incubated with heat inactivated RNPs. It was unexpected to detect the targeted product at 25ºC. This result indicates that the RNPs can reverse splice into sperm DNA at suboptimal temperature. By checking the integrated product in diluted sperm reaction DNA, it was found that the Tx1-2768a intron targeting efficiency at 25ºC was 10 times less than that at 30ºC (data not shown).

A similar temperature optimization experiment was done with Mitf-α-M-609s targetron RNPs for several reasons. First, the Tx1 gene is a multiple copy gene. The targeting efficiency cannot be accurately determined by PCR analysis of extracted sperm DNA. It could not differentiate between targeting 10 copies of Tx1 gene in one sperm from targeting 1 copy of the Tx1 gene in 10 sperm nuclei. However experiments with single copy genes do not have this problem. Second, the temperature for targeting multiple copy genes may not be the same for single copy genes. Based on the results from Tx1-2768a RNPs, sperm TPRT of the Mitf-α-M-609s targetron was tested at temperatures of 25, 27, 29, 31 and 33ºC (Figure 4.5). Perhaps due to the lower number of target sites in the genome, the intron targeted product can only be detected through nested PCR. No targeted product was detected at temperature of 25 or 27ºC. The targeted

product can only be detected at temperatures of 29ºC or above. The efficiency of targeting at 33ºC (1 in 126 sperm nuclei) is almost 10-fold that at 29ºC (1 in 1257 sperm nuclei). Compared to the Tx1-2768a intron, lowering the reaction temperature has a more dramatic effect in decreasing the target efficiency for the Mitf-α-M-609s targetron.

The above experiments assessing intron integration efficiency at different temperatures helped me determine the optimal temperature for different RNPs in transgenesis. For the Tx1-2768a intron, the intron targeting efficiency at 25ºC was 10 times lower than that at 30ºC. Based on transgenesis experiments of fertilizing untreated sperm nuclei with eggs, the percentage of healthy embryos at 25ºC was twice that at 30ºC. Balancing both considerations, an incubation temperature of 30ºC may be best for obtaining the largest number of transgenic animals.

The results in this section showed that Ll.LtrB intron integration in sperm nuclei is a temperature-dependent reaction. Within the permissible temperature range, higher temperature results in higher integration efficiency. However, to minimize the damage to sperm nuclei, incubating the sperm TPRT reaction around 30ºC to 33ºC is acceptable to ensure the efficiency of intron targeting yet mild enough for transgenesis purposes. It is critical to balance between high targeting efficiency and maintaining the integrity of sperm nuclei for transgenesis.

### 4.2.3 Nucleoplasmin in the sperm reaction

In group II intron-based *X. laevis* transgenesis, the recombinant nucleoplasmin is used to decondense sperm nuclei chromatin structure instead of *X. laevis* egg extract used in the standard transgenesis protocol (Smith et al. 2006). Nucleoplasmin is a nuclear protein that mediates the correct association of DNA with histones. During fertilization, nucleoplasmin can remodel the highly condensed paternal chromatin by replacing sperm

basic protein with histones. Each monomer of this pentameric protein consists of two domains: the core and the C-terminal tail. The core domain is responsible for nucleoplasmin oligomerization and confers extremely high stability to the protein (the denaturing temperature of the recombinant core domain at neutral pH is 120ºC). The activity of nucleoplasmin is modulated by phosphorylation (Cotten et al. 1986; Leno et al. 1996). The recombinant core can be activated through mutations that mimic phosphorylation. Such a mutant Core8D nucleoplasmin was found to dramatically decondense sperm chromatin structure at 32 μM and 64 μM (Banuelos et al. 2003).

The strong decondensation activity and extreme stability of nucleoplasmin makes it very attractive for transgenesis applications. First, it is a protein naturally involved in sperm chromatin remodeling. Therefore, it is not toxic to the sperm or embryos. Second, its extreme stability facilitates protein purification. By contrast, the decondensation activity of egg extract can vary from batch to batch due to egg quality and/or preparation procedures. The storage time of egg extract is also not as long as that of nucleoplasmin. Third, for the sperm TPRT optimization, the degree of sperm nuclei decondensation can be controlled by the amount of nucleoplasmin in the reaction, while such control is more difficult to achieve with egg extracts. Last, the purified nucleoplasmin is cleaner than egg extracts in terms of nucleases, which can inactivate the RNP by degrading the intron RNA.

Tx1-2768a intron sperm TPRT experiments followed by real-time PCR analysis showed that using 0.63 μM and 1.26 μM of nucleoplasmin in the reaction gave 23% and 16% intron integration efficiency, respectively (data not shown), suggesting that higher concentrations of nucleoplasmin can inhibit intron targeting. Similar inhibition was observed in the Mitf-α-M-609s intron targeting (data not shown). It is possible that

excess nucleoplasmin disassembles the RNP particles by binding to the nucleic acid in the RNP. Thus, for a specific intron, the optimal amount of nucleoplasmin used in the reaction needs to be determined. Too much nucleoplasmin may not only inhibit the RNP activity, but also lead to excessive decondensation that is detrimental to the sperm. Nevertheless, sufficient nucleoplasmin is needed to enable efficient intron integration.

### 4.3 TRANSGENIC TADPOLES WITH Tx1 INTRON INTEGRATION

*In situ* hybridization experiments have shown the Tx1 element is distributed over all chromosome arms within the *X. laevis* genome (Garrett and Carroll 1986). The abundance of Tx1 gene throughout the genome increases the chance of intron integration.

To obtain embryos containing targeted Tx1 integrations, a sperm TPRT reaction was done by incubating 10 μl of Tx1-2678a targetron RNPs (1.8 mg/ml lariat RNA based on $O.D._{260}$) with 4.7 μM nucleoplasmin and sperm nuclei at 30ºC for 15 min. After the reaction, the reaction mixture was diluted 30 times using sperm dilution buffer (SDB) and injected into freshly prepared *X. laevis* eggs. In one experiment, 2,200 eggs were injected and 48 viable embryos were obtained. Sixteen out of the 48 embryos developed into healthy tadpoles, which were subjected to tail-clipping for PCR analysis as shown in Figure 4.6A. The remaining 32 embryos were not healthy and were frozen after stage 22 (Nieuwkoop 1994) for the genomic DNA extraction. Three of these embryos were found to have the expected 3' junction for site-specific integration of the Tx1-2768a targetron (Figure 4.6B).

Further inverse PCR analysis showed that each of the three tadpoles contained a 5'-truncated intron inserted into a Tx1 target site. The 5' ends of truncated introns in tadpoles were indicated in the Ll.LtrB RNA secondary structure (Figure 4.10). The 5' termini of tadpole B34 and 93 are present in a single-strand loop region and the 5'

terminus of the tadpole 80 is located near a single-strand loop. This suggests that the 5' truncations could result from incomplete cDNA synthesis due to the nicked or degraded intron. It could also be due to the dissociation of RT in loop regions prior to regions of stable RNA secondary structure. Another possibility is the 5' truncations are due to degradation of the cDNA by enzymes involved in strand resection at a double-strand break that occurs during intron integration. However, from the inverse PCR, the sequences upstream of integrated introns were not the expected Tx1 gene but some unknown sequences (Figure 4.6B). It is possible that the 5' exon was resected during ligation and repair. The unknown sequences could be unsequenced intron regions in *X. laevis* genome.

A second transgenesis experiment using the Tx1-2768a intron showed the successful integration of full-length Ll.LtrB intron into *X. laevis* genome. In this experiment, the sperm TPRT reaction was done by incubating 10 μl of RNP (approximately 3 mg/ml lariat RNA based on $O.D._{260}$) with 4.7 μM nucleoplasmin and sperm nuclei at 33ºC for 15 min. After the incubation, the reaction mixture was diluted 30 times with SDB buffer and injected into 1,500 freshly prepared eggs. After two hour incubation at room temperature, 399 embryos reached the normal 4-cell stage and were sorted into fresh incubation buffer, as described in Methods. After overnight incubation, only 3 embryos were healthy at the gastrulation stage possibly due to the poor egg quality since a high proportion of control eggs injected with untreated sperm nuclei also failed to develop normally. Although the embryos were not healthy, they continued growing to hatching. Eventually, 40 embryos developed for at least 76 h post-fertilization. DNA was extracted from the individual embryo for PCR analysis. As shown in Figure 4.6C, the embryo 28 was found to have both the expected 5'- and 3'-intron integration junction

within the Tx1 gene. In a negative control, genomic DNA from an untreated tadpole did not give the same PCR bands. The identities of PCR bands corresponding to the 5'- and 3'- integration junctions were further confirmed by sequencing (Figure 4.6C).

The above transgenesis experiments demonstrated that the Ll.LtrB intron can integrate site-specifically into a desired target site in a eukaryotic genome. The results proves that: (i) sperm nuclei containing targeted group II intron modifications can fertilize eggs and develop into healthy tadpoles with the site-specific modification in sperm; (ii) transgenic frogs with a disrupted Tx1 gene are viable until the tadpole stage; (iii) both full-length or 5'-truncated introns can be integrated into the Tx1 gene.

### 4.4 SPERM TPRT REACTIONS FOR SINGLE COPY PROTEIN CODING GENES (TYR AND MITF)

The Tyr-401a targetron is designed to target exon 2 of the tyr gene in *X. laevis*. An *in vitro* sperm TPRT reaction for targetron Tyr-401a is shown in Figure 4.7A. In this reaction, I used 10 mM $Mg^{2+}$, 5 μM nucleoplasmin, 4 μl of sperm nuclei and 13.4 μl of RNP (2 mg/ml lariat RNA based on $O.D._{260}$). The reaction mixture was incubated at 33ºC for 10 min. dNTPs were then added to 0.2 mM and the reaction continued for another 20 min at 33ºC. The reaction was stopped by heating at 75ºC for 10 min. The total DNA from sperm TPRT reaction was extracted and analyzed by PCR. Due to lower targeting efficiency in single copy genes, a nested PCR was required for efficient detection of the targeted integration. Figure 4.7A shows detection of the 3' junction of the integrated intron into the tyr gene target site. No targeted product was present in a parallel sperm TPRT reaction using the same amount of heat inactivated Tyr-401a RNP, which demonstrated that the targeted product arose from Tyr-401a intron reverse splicing activity. The other bands in the figures are unspecific PCR products and verified by the sequencing analysis. Further PCR analysis using serial diluted sperm DNA from the

reaction revealed that the targeting efficiency of Tyr-401a targetron was 1 in 301 sperm nuclei (0.3%, data not shown).

The Mitf-α-M-609s targetron is designed to target exon 6 of the mitf gene in *X. laevis*. An *in vitro* sperm TPRT reaction for targetron Mitf-α-M-609s is shown in Figure 4.7B. Targeting with Mitf-α-M-609s was achieved by incubating 14 μl of RNPs (1.5 mg/ml lariat RNA based on $O.D._{260}$) with 10 mM $Mg^{2+}$, 4 μl of sperm nuclei, 2.2 μM of nucleoplasmin at 30ºC for 5 min. dNTPs were then added to 0.2 mM, and the reaction was continued for another 10 min. The sperm TPRT reaction was then stopped by heat inactivation. The total nucleic acid was extracted and the 3'-integration junction was detected by nested PCR as shown in Figure 4.7B. After serially diluting the sperm TPRT DNA, the targeted product of the Mitf-α-M-609s intron was detected in 508 sperm nuclei (0.2% integration efficiency). The PCR product, indicated by arrow, is the expected junction product. Other bands were checked by sequencing and confirmed to be unspecific PCR products. The same intron targeting efficiency was subsequently confirmed by real-time PCR (data not shown).

### 4.5 STUDY OF LINEAR GROUP II INTRON LL.LTRB

Like spliceosomal introns, most group II introns splice via two sequential transesterification reactions that result in the formation of an intron lariat RNA (Pyle and Lambowitz 2006). For mobile group II introns, the splicing reactions are assisted by the intron-encoded RT, which binds specifically to the intron RNA to stabilize the catalytically active RNA structure and then remains bound to the excised intron lariat RNA in RNPs that promote intron mobility (Saldanha et al. 1999; Matsuura et al. 2001).

Some group II introns differ in using a variation of the splicing mechanism in which cleavage at the 5'-splice site occurs by hydrolysis rather than branch-point

formation, leading to the production of linear rather than lariat intron RNA. Such "hydrolytic splicing" was first observed in group II intron self-splicing reactions under non-physiological conditions (*i.e.*, high salt and/or $Mg^{2+}$ concentrations) (Van Der Veen et al. 1987; Jarrell et al. 1988; Daniels et al. 1996) and shown to occur *in vivo* by using branch-point mutants of the yeast mt group II intron aI5γ intron (Podar et al. 1998). Some naturally occurring group II introns lack a bulged branch-point nucleotide in domain VI, and at least one of these, a plant chloroplast *trnV* group II intron, splices *in vivo* by a mechanism that produces linear RNA without detectable amounts of lariat RNA (Vogel and Borner 2002). In addition to hydrolytic splicing, linear group II intron RNA can also be generated from excised lariat RNAs by the action of debranching enzyme, which is used in eukaryotes for the debranching and turnover of excised spliceosomal introns (Green 1986).

Linear group II intron RNAs cannot carry out both steps of reverse splicing and were thus thought to be immobile. However, the linear intron RNAs can carry out the first step of reverse splicing into RNA or DNA sites, thereby ligating the 3' end of the RNA to the 5' end of the 3' exon (Morl and Schmelzer 1990; Mastroianni et al. 2008; Roitzsch and Pyle 2009). This partial reverse splicing reaction could potentially be used for mobility. If partial reverse splicing occurs into RNA, then the recombined RNA would need to be reverse transcribed and the resulting cDNA integrated into the host genome by recombination, while if partial reverse splicing occurs into DNA, the attached intron RNA could be reverse transcribed and the intron cDNA integrated by DNA repair. The latter pathway is most likely for group II introns that encode IEPs because association with the IEP in RNPs strongly biases the intron RNA to reverse splice into DNA rather than RNA sites (Zimmerly et al. 1995a).

Recently, we investigated whether *Lactococcus lactis* Ll.LtrB intron RNPs reconstituted with purified IEP and linear intron RNA (Lin RNPs) could use their reverse splicing and second-strand cleavage activities to introduce a double-strand break at a DNA target site (Mastroianni et al. 2008). The introduction of a recombinogenic double-strand break by protein endonucleases, such as a Zn-finger nucleases or meganucleases, is a favored mode of gene targeting in higher organisms (Porteus and Carroll 2005), and we reasoned that group II introns might be advantageous for this approach, as they can be retargeted readily to introduce the double-strand break at desired locations simply by modifying the EBS and δ sequences in the intron RNA (Perutka et al. 2004).

### 4.5.1 Biochemical activities of linear group II intron Ll.LtrB RNPs

I carried out biochemical assays to test whether Ll.LtrB Lin RNPs could efficiently carry out the first step of reverse spicing into DNA target sites, En domain cleavage and reverse transcription of the attached linear intron RNA (Mastroianni et al. 2008). Figure 4.8A shows an experiment in which linear and lariat intron RNPs were incubated *in vitro* with an internally labeled double-stranded DNA oligonucleotide substrate containing the Ll.LtrB-insertion site, and the products were analyzed by electrophoresis in a denaturing polyacrylamide gel. As found previously (San Filippo and Lambowitz 2002), the RNPs containing lariat RNA gave products resulting from both full and partial reverse splicing of the intron RNA into the top strand, along with two closely spaced fragments resulting from IEP cleavage of the bottom strand (Figure 4.8A, lane 3). By contrast, the RNPs containing linear intron RNA did not yield the same fully reverse-spliced product and intermediate containing attached lariat RNA as did lariat RNPs, but did cleave the top strand by partial reverse splicing, resulting in a small band corresponding to the 5'-exon (5' top) and a high-molecular weight product containing

linear intron RNA attached to the 5' end of the 3'-exon (Figure 4.8A, lane 2). The RNPs containing linear intron RNA also carried out IEP cleavage of the bottom strand, producing the same two closely spaced fragments as lariat RNPs (5' and 3' bottom; Figure 4.8A, lane 2).

Figure 4.8B and C show additional reactions in which RNPs containing linear or lariat intron RNA were incubated with DNA substrates labeled at the 5' end of the top or bottom strand, respectively, in the presence of dNTPs to support target-DNA-primed reverse transcription. The reactions with the 5' top-strand labeled substrate confirm that RNPs reconstituted with linear intron RNA cleave the top strand and show that the reaction can occur as efficiently as cleavage by lariat RNA (Figure 4.8B, lanes 1 and 2). The reactions with a 5'-labeled bottom strand show that the cleaved bottom strand can be used as a primer for reverse transcription of the linear intron, yielding higher molecular weight product with different lengths of attached cDNA (Figure 4.8C, lanes 1 and 2). The gels also show that cDNA synthesis increases the proportion of fully reverse-spliced product obtained with lariat RNPs, presumably by pulling the unfavorable equilibrium for full reverse splicing toward completion (Figure 4.8B, lane 2) (Aizawa et al. 2003), and that lariat or linear intron RNPs reconstituted with mutant LtrA protein that lacks RT activity can still carry out reverse splicing and top- and bottom-strand cleavage even though they cannot synthesize cDNA (Figure 4.8A, lanes 4 and 5, and Figure 4.8B and C, lanes 3).

Importantly for gene targeting, the biochemical experiments show that group II intron RNPs reconstituted with linear intron RNA can introduce a double-strand break resulting from top and bottom strand cleavage at the DNA target site. As predicted from the biochemical experiments, *X. laevis* oocyte nuclei microinjection experiments showed

that RNPs containing linear Ll.LtrB intron RNA could be used to introduce a double-strand break at a plasmid target site, thereby stimulating gene targeting by homologous recombination with a coinjected DNA fragment (Mastroianni et al. 2008).

### 4.5.2 Retrohoming of group II intron Lin RNPs in *X. laevis* oocyte nuclei

In addition to demonstrating the potential utility of group II introns for gene targeting by double-strand-break-stimulated homologous recombination, the above experiments raised the possibility that group II intron Lin RNPs might also be able to retrohome *in vivo* by a mechanism involving partial reverse splicing of the linear intron RNA into the DNA target site followed by reverse transcription of the attached RNA, provided the resulting intron cDNA could be integrated into the recipient DNA by DNA repair (Mastroianni et al. 2008).

To test the retrohoming of Ll.LtrB Lin RNPs, we used a plasmid-based assay first developed for bacteria (Guo et al. 2000) and later adapted to test retrohoming of Ll.LtrB lariat RNPs (Lar RNPs) microinjected into *X. laevis* oocyte nuclei (Mastroianni et al. 2008). In this assay, an Ll.LtrB intron with a phage T7 promoter inserted near its 3' end integrates into a target site cloned in a recipient plasmid upstream of the promoterless $tet^R$ gene, thereby activating that gene. For assays in *X. laevis* oocytes, we use the recipient plasmid pBRR3-ltrB, which carries the Ll.LtrB target site/$tet^R$ gene cassette and a separate $amp^R$ marker, and RNPs reconstituted with purified IEP and *in vitro*-synthesized 0.9-kb Ll.LtrB-ΔORF intron RNA carrying a T7 promoter. The recipient plasmid was injected first into the nucleus of stage 6 oocytes, followed within 1 min by Ll.LtrB Lar or Lin RNPs. After incubation for 2 h at 25°C, nucleic acids were isolated and transformed into *E. coli* HMS174(DE3), which expresses phage T7 RNA polymerase. The cells were then plated on LB medium containing ampicillin plus tetracycline or ampicillin alone,

and the integration efficiency (%) was quantified as the ratio of $Tet^R$ + $Amp^R$/$Amp^R$ colonies.

First, we used this assay to compare the retrohoming efficiencies of Lar and Lin RNPs. In initial experiments, we tested two forms of the linear intron RNA, one whose 5' end corresponds precisely to that of the Ll.LtrB intron (Lin) and the other with two extra 5' G residues (GG-Lin), enabling more efficient transcription by phage T3 RNA polymerase (Bailey et al. 1983). These extra 5' residues are not expected to affect the ability of the GG-Lin RNA to carry out the first step of reverse splicing, nor to serve as a 5'-exon for RNA splicing to produce lariat RNA (Jacquier and Rosbash 1986), which was confirmed by *in vitro* splicing assays (data not shown).

Table 4.1 summarizes the results of four separate experiments, with ten injected oocytes pooled for each condition in each experiment. The retrohoming efficiency of Ll.LtrB Lar RNPs in these experiments was 8-22%, in agreement with previous results (Mastroianni et al. 2008), while the retrohoming efficiencies of the Lin and GG-Lin RNPs were 0.004-0.04%, readily detectable in these assays but more than 500- to 2,000-fold lower than for Lar RNPs. Controls showed no retrohoming for GG-Lin RNPs reconstituted with an RT-deficient mutant LtrA protein (DD-; Table 4.1), as shown previously for lariat-containing RNPs (Mastroianni et al. 2008). Thus, retrohoming of both Lin and Lar RNPs in *X. laevis* oocyte nuclei requires the RT activity of the group II IEP.

The reverse splicing activity of linear intron RNA depends critically upon the RNA having the correct 3' nucleotide residues. This dependence is problematic for RNAs synthesized with phage T7 RNA polymerase, which adds extra non-templated nucleotide residues to the ends of run-off transcripts (Milligan et al. 1987), necessitating the use of

special measures, such as transcription from DNA templates with 2'-OMe nucleotide residues at the penultimate or last two nucleotide, to reduce 3' end heterogeneity (Kao et al. 2001; Roitzsch and Pyle 2009). For our experiments, linear Ll.LtrB intron RNAs were transcribed with phage T3 RNA polymerase, which differs from T7 RNA polymerase in giving run-off transcripts with homogenous 3' ends (Majumder et al. 1979). As expected, GG-Lin intron RNAs transcribed with T3 RNA polymerase from DNA templates containing 2'-OMe nucleotide residues at the last two positions did not significantly increase retrohoming frequencies (Table 4.1).

Finally, we confirmed that GG-Lin RNPs with the branch-point A-residue deleted (GG-Lin-ΔA2486) still retained substantial retrohoming activity, albeit two- to four-fold lower than that for wild-type GG-Lin RNPs assayed in parallel (0.001-0.002% and 0.002-0.008%, respectively; Table 4.1). The insertion of the ΔA2486 intron was confirmed by sequencing through the branch-point region. These findings indicate that retrohoming of GG-Lin RNPs is not dependent upon branching at the normal site. It is not clear whether the somewhat decreased retrohoming efficiencies of GG-Lin-ΔA2486 RNA are significant and if so, whether they reflect differences in the quality of the RNP preparations or somewhat decreased reverse splicing activity due to the effect of deleting the branch-point nucleotide on the intron RNA structure.

### 4.5.3 Sequences of integration junctions for retrohoming of Ll.LtrB Lin RNPs in *X. laevis* oocyte nuclei

To confirm retrohoming of the linear intron RNA, we carried out PCR of randomly selected $Tet^R$ + $Amp^R$ colonies, using primers flanking the intron-insertion site in the recipient plasmid. For Lar RNPs, the colony PCR gave products of the size expected for insertion of the full-length Ll.LtrB-ΔORF intron for all retrohoming events

analyzed, while for Lin RNPs, the colony PCR gave some products of the size expected for insertion of the full-length intron, along with smaller products (Figure 4.9A).

Sequencing of the 5'- and 3'-integration junctions for Lar RNPs showed precise insertion of the full-length intron at the target site, as expected for full reverse splicing (> 23 retrohoming products analyzed in this and previous work (Mastroianni et al. 2008)). For Lin RNPs. sequencing showed that all 24 retrohoming products analyzed had the 3'-junction sequences, expected for the first step of reverse splicing into the DNA target site. However, only one event analyzed for Lin RNPs gave insertion of full-length intron with accurate 5'- junction sequences (Figure 4.9B). The remainder had 5'-junctions sequences with some combination of: (i) resection of 5'-exon sequences; (ii) insertion of 5'-truncated introns; and/or (iii) insertion of extra nucleotide residues at the integration junction, presumably reflecting errors during the DNA repair process used for ligation of the intron cDNA to the 5' exon (Figure 4.9B). Similar 5'- and 3'-junction sequences were found for retrohoming products of the GG-Lin RNA (data not shown). A number of the junction sequences, particularly those with large 5' intron truncations, showed microhomologies of 1-5 nt residues between the 5' end of the intron and the ligated exon sequences (Figure 4.9B). The characteristics of the 5' junctions for retrohoming of Lin and GG-Lin RNPs are similar to those for ligation junctions formed during DNA repair by non-homologous end-joining (NHEJ) (Hagmann et al. 1998; Weterings and Van Gent 2004).

The fully analyze the spectrum of retrohoming products for Lin RNPs and confirm that cDNA ligation occurred in the eukaryotic host, we carried out direct PCR of 5'- and 3'- integration junctions in nucleic acids extracted from *X. laevis* oocyte nuclei without transformation into *E. coli* (Figure 4.10). As expected from the colony PCR

results, the 5'-junction PCR for Lin and GG-Lin RNPs showed some products of the same size as those for Lar RNPs, along with multiple smaller products (Figure 4.10A). Sequencing of cloned PCR products confirmed that the larger products (PCR band a) had insertions of full-length or almost full-length introns, while products of decreasing size (PCR bands b and c) had insertions of introns with progressively longer 5' intron truncations (Figure 4.10B). The direct PCR permitted the recovery of introns with longer 5' truncations that delete the T7 promoter required for the genetic selection in the bacteria. The 5'-junction sequences also showed variably sized 5'-exon resections insertion of extra nucleotide residues at the integration junction, and microhomologies in some cases, similar to the junction sequences analyzed by colony PCR (Figure 4.10B). By contrast to the 5' junction, the 3'-junction PCR for retrohoming products of Lin and GG-Lin RNPs gave a single prominent band of the same size as that for Lar RNPs (Figure 4.10), and sequencing showed the expected junction corresponding to the 3' end of the intron inserted precisely at the target site by partial reverse splicing in all cases (data not shown).

### 4.5.4 Model for retrohoming of linear group II intron RNA

Our results indicate that RNPs containing linear group II intron RNA retrohome *in vivo* by the mechanism shown in Figure 4.11. The pathway begins with the linear intron RNA catalyzing the first step of reverse splicing into a DNA target site, resulting in the ligation of the 3' end of the intron to the 5' end of the 3' exon DNA. The associated IEP then uses its En domain to cleave the opposite strand and synthesizes a cDNA copy of the linear intron RNA. These initial steps were demonstrated previously for purified GG-Lin RNPs *in vitro* (Mastroianni et al. 2008). A key step shown here to required to retrohoming *in vivo* is the ligation of the free end of the intron cDNA to the

upstream exon DNA. This step occasionally leads to the precise insertion of the intron, just as for retrohoming of lariat RNA, but more frequently occurs with loss of 5'-exon or 5'- intron sequences and insertion of additional nucleotide residues at the junction. As we observed no duplication of target site sequences resulting from the initial staggered double-strand break made by the RNPs, we infer that the single-stranded 5' overhang attached to the 5' exon on the bottom strand must be resected prior to cDNA ligation. The retrohoming of linear intron RNA is presumably completed like retrohoming of lariat RNA by the degradation or displacement of the intron RNA template strand followed by second-strand DNA synthesis and the sealing of nicks by host DNA polymerases and ligases (Smith et al. 2005). The characteristics of the cDNA ligation step, in particular the 5'-exon resection and the insertion of extra nucleotide residues at the ligation junction, suggest that it occurs by non-homologous end joining (Hagmann et al. 1998; Weterings and Van Gent 2004).

The insertion of 5'- truncated introns likely results from degradation of the linear intron RNA rather than abortive cDNA synthesis, as such 5' truncations are not observed for retrohoming via lariat RNA, which is identical to linear intron RNA except that its 5' end is protected first by the 2'-5' linkage and then by ligation to the 5' exon after the second-step of reverse splicing. The majority of 5' truncated ends (60-70%) fall in single-stranded loop regions, which are expected to be particularly sensitive to ribonucleases, with some loops appearing to be hot spots for the generation of truncated cDNAs (Figure 4.12).

### 4.6 Ll.LtrB intron with GFP-RAM (Retrotransposition-activated marker)

To apply group II intron-based gene targeting in sperm nuclei to protein-coding genes, it is desirable to incorporate a genetic marker into the intron to facilitate the screening of embryos for intron integrations. Two problems were: (i) inserted genetic markers often decrease the efficiency of group II intron RNA splicing and reverse splicing reactions, likely by impeding RNA folding, leading to substantially decreased DNA integration frequencies (Plante and Cousineau 2006), and (ii) RNPs reconstituted with *in vitro* transcribed intron RNA generally contain small amounts of residual template DNA. This residual DNA is difficult to remove even with extensive DNase treatment and could integrate randomly producing a background of embryos expressing the marker from the integrated DNA.

To address these problems, a GFP-RAM marker that functions efficiently when inserted into intron DIV was developed (Figure 4.13A). This marker consists of a GFP coding sequence, which has been optimized for function in *X. laevis* embryos (Zernicka-Goetz et al. 1996), transcribed from a CMV promoter and inserted in the reverse orientation to the intron transcription, but interrupted by an efficiently self-splicing group I intron (phage T4 *td* intron) in the forward orientation. After *in vitro* transcription, the *td* group I intron is excised by self-splicing to yield Ll.LtrB-ΔORF intron RNA with an intact GFP marker for RNP reconstitution. To make this barrier even more stringent, I used a mutant *td* intron (G67A mutation in P5 element) that requires higher $Mg^{2+}$ concentrations (8 mM) for self-splicing, ensuring that it cannot be removed by self-splicing under *in vivo* conditions (Mohr et al. 1992). Because of the presence of the *td* intron, green fluorescent embryos cannot arise from random integration of residual DNA template. For linear introns, the RNA can be used directly for RNPs reconstitution after

*in vitro* splicing of the *td* intron. For lariat introns, an additional Ll.LtrB self-splicing step is required to form a lariat RNA with an intact GFP marker in DIV.

### 4.6.1 The splicing efficiency and *in vitro* TPRT activity of Ll.LtrB intron with GFP-RAM

Inserting genetic markers into group II intron DIV often decreases the efficiency of group II intron RNA splicing and reverse splicing (Plante and Cousineau 2006). Here, I assessed the splicing efficiency of the Ll.LtrB intron with the GFP-RAM by electrophoresis in a denaturing 4% polyacrylamide gel and real-time PCR.

The Ll.LtrB intron with the GFP-RAM was found to splice efficiently based on electrophoresis in a denaturing 4% polyacrylamide gel analysis (Figure 4.13B). In this experiment, a precursor RNA containing the Ll.LtrB intron with the GFP-RAM was transcribed from the linearized plasmid pACD-GFP-RAM. Two splicing steps were carried out, one to splice the *td* intron, and the other to form the lariat Ll.LtrB intron with intact GFP. The initial transcript and spliced product were loaded in parallel lanes for comparison. The major band in the transcript lane was the precursor RNA. Other smaller bands arose from low level *td* splicing which occurred during the transcription. Compared to the input lane, about half of precursor RNA converted into the lariat form, corresponding to the top band in the "Spliced" lane. Some precursor was converted into a smaller linear form. This linear form could be linear intron RNA with or without the *td* intron, which are indistinguishable on the gel. The result was repeatable and demonstrated that the Ll.LtrB with GFP-RAM marker can splice efficiently *in vitro*.

However, the splicing efficiency of the *td* intron could not be determined from the gel. It was checked by a one step RT real-time PCR using *power* SYBR Green RNA-to-$C_T$ *1-Step* Kit (Applied Biosystems). To compare the amount of RNA with or without the *td* intron, two primers were designed to specifically amplify the designated products (blue

arrows in Figure 4.13A). The forward primer for precursor RNA ($P_{pre}$) anneals to the 3' end of the *td* intron, ending 2-nt upstream of the 3'-splice site, while the forward primer for the spliced product ($P_{sp}$) spans the splice junction, with 24-nt in the 5'-exon and 4-nt in the 3'-exon, thereby ensuring that it will only amplify the ligated exons and not precursor RNA. The common reverse primer ($P_{rev}$) is complementary to a sequence 127-nt downstream of the splice site. The specificity of the primers for their intended amplicons was checked by RT-PCR. Both sets of primers only amplified the designed products (data not shown). Therefore the *td* splicing efficiency can be determined by the ratio of spliced RNA/ (precursor RNA + spliced RNA) = $2^{\Delta Ct}/ (1+2^{\Delta Ct})$, in which $\Delta Ct = Ct_{precursor}-Ct_{spliced}$. Based on the real-time PCR results, the *td* splicing efficiency is about 70% (Figure 4.13C).

Furthermore, sperm TPRT reactions analyzed by real-time PCR to quantitate the 3'-integration junction indicated that Tx1-2768a and Tx1-2768a GFP-RAM RNP are about equally efficient for site-specific DNA integration (data not shown). *In vitro* TPRT assay showed that Mitf-α-M-609s GFP-RAM RNPs can actively reverse splice into the DNA target site in the plasmid, although the reverse splicing activity was lower than that of Mitf-α-M-609s RNP in the same experiment (data not shown). Together, these findings indicate that although the Ll.LtrB-GFP-RAM construct may have somewhat decreased splicing and reverse splicing activity as compared to Ll.LtrB-ΔORF construct, it still retains relatively high activity in both the forward and reverse splicing reactions.

### 4.6.2 Site-specific gene disruptions using an Ll.LtrB intron with a GFP marker

A construct named pACD-Mod4-GFP was made during the construction of pACD-GFP-RAM. It contained an Ll.LtrB intron with the GFP cassette but no *td* intron. This construct was used to make Tx1-2768a RNPs for transgenesis. In this experiment,

the sperm TPRT reaction contained 10 mM $Mg^{2+}$, 0.2 mM dNTP, 15.5 μl of RNP (1.5 mg/ml lariat RNA based on $O.D._{260}$) and permeabilized sperm nuclei in SDB buffer. Nucleoplasmin was not used to increase the chance of obtaining healthy tadpoles. The reaction mixture was incubated at 30ºC for 10 min and then diluted 30 times using SDB with 5 mM $Mg^{2+}$ for transgenesis. A total of 2,720 eggs (~160 eggs per plate, 17 plates total) were injected with reacted sperm nuclei. After incubating at 18ºC for 3 h, 11.5% ~ 21.2% of the embryos in each plate developed normally to the 4-cell stage and 58% ~ 80.2% of these healthy 4-cell embryos in each plate developed to stage 19 (Nieuwkoop 1994). After 8 days, 61 embryos developed into healthy swimming tadpoles. Although GFP expression was not detected in any of the embryos (81 embryos including the 61 healthy tadpoles and 20 unhealthy ones), they were checked by PCR for the 3'-intron integration junction. Eight embryos showed to have the correct 3' intron integration junction (data not shown). The fact that none of them showed the expression of GFP throughout the development could reflect the integration of 5'-truncated introns that lack part or all of the GFP marker. Some tadpoles grew in the lab for more than three months, confirming that the disruption of the Tx1 gene did not affect embryo development.

### 4.7 LINEAR LL.LTRB-GFP-RAM INTRON LIBRARY WITH RANDOM EBS 1/2 AND δ SEQUENCES

In previous experiments with *E. coli*, a library of Ll.LtrB-ΔORF introns with randomized EBS and δ sequences integrated at sites throughout the *E. coli* genome, generating a gene disruption library (Zhong et al. 2003). The integrated introns, which contain EBS and δ complementary to the IBS and δ' sequences at their insertion sites, could be recovered by PCR and in many cases, used directly without further modification to generate the desired single disruptions (Yao et al. 2005). This approach has the

advantages of identifying accessible target sites and efficient introns without resorting to the computer algorithm.

A similar procedure was applied in *X. laevis* by using an Ll.LtrB-GFP-RAM intron library containing linear intron RNA with randomized EBS1/2 and δ sequences. A lariat GFP-RAM RNP library was difficult to obtain because of the low splicing efficiency of precursor RNA with random IBS and EBS sequences (Jacquier and Michel 1990), but the discovery that linear intron can promote intron integration by reverse splicing and nonhomologous end-joining in *X. laevis* oocytes resolved this problem. By using the linear Ll.LtrB intron, the transcribed RNA with randomized EBS1/2 and δ sequences can be used directly for *in vitro* RNP reconstitution after the *td* splicing. The GFP expression was used to screen the embryos for intron integrations (Figure 4.14A).

In the first experiment, 13 μl of GFP-RAM RNP library (linear RNA concentration 2.2 mg/ml based on $O.D._{260}$) was used in a sperm reaction with 10 mM $Mg^{2+}$, 2 mM dNTP, and 2.5 μM nucleoplasmin. The reaction was incubated at 30°C for 10 min and the modified sperm nuclei were diluted 30 times in 1×SDB with 5 mM $Mg^{2+}$, and then used directly for fertilization. After incubation at 18°C, embryos were checked under the fluorescent stereoscope to check GFP expression. Figure 4.14B showed a field view of embryos that developed for three days at 18°C. Under the UV light, around 5% of the embryos in the field show GFP expression. The individual $GFP^+$ embryos were frozen and genomic DNA was extracted for PCR analysis.

To identify the integrated introns and their insertion sites in the $GFP^+$ embryos, TAIL PCR (Thermal Asymmetric Interlaced PCR) was used. This allowed amplification of the 5' part of Ll.LtrB intron, including the modified EBS1/2 and δ sequences, along with the *X. laevis* genomic DNA sequences upstream of integrated intron. After the TAIL

PCR, a sequence specific primer annealing to the genomic DNA upstream of the integrated intron was designed and used for confirmation the 5'-intron integration junction by direct PCR amplification.

Table 4.2 summarizes the confirmed 5'-integration junction sequences for this experiment (Exp 1). The results show that the introns integrated into both protein coding regions and unsequenced regions of the *X. laevis* genome (denoted "Unknown"). Both truncated and full intron integrations were found in this experiment. Based on the sequences of 5'-integration junctions, eight of the integrations had full-length Ll.LtrB introns, while the other four had 5' truncated introns. Some embryos have more than one intron integration, *e.g.*, introns 12, 13 and 14 were recovered from one embryo. Further, the same intron integration was found in several different embryos (note multiple occurrences for introns 9 and 13 in Exp. 1, Table 4.2). These results suggest that some introns may be very active and/or that some target sites may be hot spots for intron integration. The introns 12 and 14 integrated into genome at the same site with identical 5' truncations (Table 4.2 Exp.1). While I cannot exclude that this is due to PCR contamination, such contamination was not detected in the PCR water control done in parallel. Because the EBS and δ sequences are deleted, we cannot deduce whether these are the same intron or different introns that coincidently integrated with the same 5' truncation. Some integrated introns have similar EBS1, EBS2 or δ sequences and integrated into the same site in the genome – *e.g.*, introns 22 and 27, which have exactly the same δ and EBS1 sequences and EBS2 sequences differing by one T residue.

The second experiment was carried out using the same transgenesis conditions as the first one, but with a different preparation of the GFP-RAM RNP library (linear RNA concentration 9.5 mg/ml based on $O.D._{260}$). After screening for embryos with GFP

expression and confirming intron integration by PCR, three different introns were found. Interestingly intron 1 in Exp. 2 integrated into four different embryos. Sequencing results revealed that it had wild-type Ll.LtrB intron sequences for EBS1/2 and δ.

Most of the embryos with GFP fluorescence in the previous trials showed developmental abnormities (Figure 4.14B). As I found in the course of this work that using less nucleoplasmin in the sperm TPRT reaction improves the frequency of generating healthy embryos, I carried out a third transgenesis with the same linear Ll.LtrB-GFP-RAM library as in Exp. 2 but less nucleoplasmin. In this experiment, the sperm reaction was done with 0.75 μM of nucleoplasmin and 8 μl of GFP-RAM RNP library (linear RNA concentration 9.5 mg/ml based on $O.D._{260}$). A total of 1,440 eggs were injected, and 9 embryos developed into tadpoles, four of which expressed GFP (Figure 4.14C). Genomic DNA was isolated from tail clippings of these tadpoles for DNA analysis. The TAIL PCR detected introns integrated into these tadpoles. The intron EBS1, 2 and δ sequences and flanking *X. laevis* genomic sequences are summarized in Table 4.2 Exp.3. Tadpole A and I (Figure 4.14C) were found to have the same intron (intron 3 in Exp.3, Table 4.2). However due to difficulties in this TAIL PCR, the genomic sequences upstream of intron were not readable. It is possible that some secondary structure impedes the progression of the DNA polymerase used in the PCR. Notably, intron 3 in Exp. 3 is the same as intron 1 in Exp. 2, with wild type EBS1/2 and δ sequences (Table 4.2). The TAIL PCR suggests that it integrated at the same site in the Cripto-1 gene, but this could not be confirmed by direct PCR due to difficulty with the PCR reactions. Tadpole H (Figure 4.14C) was found to have intron 2 in Exp. 3. The same intron sequence and integration site were also found in Exp. 2 (intron 6). The tadpole F (Figure 4.14C) contains intron 3. Sequencing result showed the intron integrated into an

unsequenced region in the genome. This transgenic experiment demonstrated that by using less nucleoplasmin, healthy tadpoles with random intron integrations can be obtained.

The above experiments showed that the Ll.LtrB intron with randomized EBS1/2 and δ sequences can randomly integrate into *X. laevis* genome, protein-encoding genes or non protein-encoding regions, and that some of the resulting embryos containing targeted intron integrations can develop into healthy tadpoles. This strategy can potentially be used for forward mutagenesis method in *X. laevis*.

## 4.8 DISCUSSION

### 4.8.1 Intron target sites in *Xenopus*

In this chapter, I developed a new method for group II intron-based gene targeting via site-specific DNA modification in *X. laevis* sperm nuclei. Potential target sites were identified in four genes (Tx1, 28s, mitf and tyr). The Ll.LtrB intron was retargeted to insert into those sites in sperm nuclei. For the targetron Tx1-2768a, the integration efficiency determined by real-time PCR was about 12% and was efficient enough to readily generate embryos containing the targeted modification following *in vitro* fertilization. For the single copy genes, mitf and tyr, the integration efficiencies were 0.3% and 0.2% respectively.

Recent findings showed that chromatin structure impedes intron targeting in *X. laevis* oocyte nuclei (Mastroianni et al. 2008), and this also appears to be the case in sperm nuclei, as judged by higher targeting efficiency found for Tx1-2786a RNPs in extracted sperm DNA than in sperm nuclei (Figure 4.4B). The use of strategies employing an Ll.LtrB intron with randomized EBS and δ sequences developed in section 4.7 should provide additional information about group II intron targeting preference in

eukaryotic cells, as well as identify accessible target sites that are not blocked by chromatin structure.

### 4.8.2 Sperm reaction conditions

The research described in this chapter identifies several factors that are critical for group II intron gene targeting in *X. laevis* sperm nuclei. First, the amount of nucleoplasmin used in the reaction needs to be sufficient to decondense the chromatin structure of the target site, but not so great as to affect the fertilization activity of the sperm nuclei or cause abnormalities during embryo development. In the absence of group II intron RNPs, nucleoplasmin in the concentration range of 0 to 10 μM had no effect on the fertilization, but concentrations higher than 10 μM caused developmental abnormalities at the beginning of gastrulation or later stages (unpublished data). Since the nucleoplasmin is an acidic protein, it may have relatively high binding affinity for exposed intron RNA in the RNPs. Data from Mitf-α-M-609s intron targeting showed that too much nucleoplasmin can inhibit the intron targeting activity (data not shown), suggesting that an optimal balance of nucleoplasmin and RNP concentrations may be required.

In addition to the amount of nucleoplasmin, the activity of RNPs used in the sperm TPRT reaction needs to be calibrated *in vitro* using the TPRT assay. Sperm TPRT assays with the Tx1-2768a intron targeting showed that targeting efficiency increased with increasing RNP concentrations (Figure 4.4). However, the highest amount of RNPs tested was 60 μg in a 27μl sperm reaction.

Finally, my results suggest that optimal temperatures for intron targeting in sperm nuclei are between 30ºC and 33ºC. Lower temperatures result in decreased intron integration efficiency, while higher temperatures are detrimental to the sperm, with

temperatures higher than 33ºC dramatically decreasing the fertilization rate during transgenesis.

### 4.8.3 Reactions of linear Ll.LtrB intron

In this chapter, the reverse splicing activity of RNPs reconstituted with linear Ll.LtrB intron RNP was characterized both *in vitro* and *in vivo*. *In vitro* biochemical assays demonstrated that such linear RNPs can recognize DNA target sites, carry out the first step of reverse splicing, resulting in attachment of linear intron RNA to the 3' exon DNA, cleave the bottom strand, and carry out target DNA-primed reverse transcription of the attached intron RNA. Although linear RNPs cannot carry out full reverse splicing, they could, surprisingly, still integrate at reduced efficiency into DNA target sites in *X. laevis* oocyte nuclei *in vivo* experiments. Analysis of the junction sequences from such experiments suggests that the integration ("retrohoming") mechanism for linear RNPs involves partial reverse splicing followed by reverse transcription of the attached intron RNA, followed by the use of host factors to ligate the 5'-end of the intron cDNA to the 5'-exon DNA. Host factors also presumably remove the intron RNA template strand and carry out second-strand DNA synthesis. Although some retrohoming events for linear intron RNPs precisely integrate the full-length or almost full-length intron, others have insertions, deletions, and/or mutations at the 5'-integration junction, as expected for inaccurate nonhomologous end joining. Additionally, retrohoming events with linear intron RNPs have a tendency to integrate 5'-truncated introns, probably because the unprotected 5'-end of linear intron RNA is more accessible to ribonucleases *in vivo*. In addition to revealing a novel retrohoming mechanism that might be used by group II introns that are excised as linear rather than lariat RNAs, these findings suggest that RNPs containing linear intron RNA might be useful for gene targeting, not only for

introducing a recombinogenic double-strand break, but also for site-specific integration into DNA target sites.

### 4.8.4 Random integrations with a linear Ll.LtrB-GFP-RAM intron library

The Ll.LtrB-GFP-RAM intron library with randomized EBS1/2 and δ sequences developed here may be useful as a tool for forward genetic screens in *X. laevis*. The disrupted genes with specific GFP expression patterns can be identified by TAIL-PCR and integrated introns can be isolated by PCR and potentially used to obtain the desired single disruptants. Further experiments with this library may also provide information about group II intron targeting rules in eukaryotic cells, as well as identify validated target sites that are accessible in *Xenopus* sperm nuclei. Importantly, the GFP-RAM library RNPs contain linear intron RNAs, which can be obtained readily by *in vitro* transcription without the RNA splicing step required to generate lariat RNAs. In addition to being easier to prepare, the base-pairing rules obtained by analysis of integration events with the linear intron RNA library will pertain exclusively to the reverse splicing reaction without potential bias for base-pairing interactions required for forward splicing needed to generate lariat RNA, as is the case for current gene targeting rules determined in bacteria (Perutka et al. 2004). The procedures developed here for group II intron gene targeting via sperm DNA modification should be generally applicable to *X. tropicalis* and potentially to other organisms.

## 4.9 METHODS

### 4.9.1 Recombinant plasmids

pACD2 contains the 0.9-kb Ll.LtrB-ΔORF intron and flanking exons cloned downstream of a T7/lac promoter in a pACYC184-based vector with a $cam^R$ gene, and

has the LtrA ORF cloned downstream of the 3' exon. The LtrA ORF was removed from the intron DIV and replaced by a phage T7 promoter (Guo et al. 2000; Karberg et al. 2001). pACD3 is a derivative of pACD2 without the T7 promoter in intron DIV (Perutka et al. 2004).

pMod4 is related to pACD3 but contains an alternate ΔORF construction based on the predicted intron secondary structure, with MluI and EcoRV sites introduced at the site of the ORF deletion site to facilitate the insertion of selectable markers. To make the intended deletion and add the two restriction sites, the 5' and 3' parts of the intron with an overlapping sequence were amplified separately by PCR of pLE12, which contains the full-length Ll.LtrB intron (Matsuura et al. 1997). Primers Ex1F (5'-AGCTTgaattcCTTAGAGAAAAATAATGCGGTGC-3') and Mod4R (5'-CGTGTGGGCGATAAAacgcgtGTCTTGTAAAAACTTCGTCT-3') were used to amplify the 5' part of Ll.LtrB intron, and primers Mod4F (5'-TTTTATCGCCCACACgatatcTCCCGTTAAAAGCTAAATGT-3') and Ex2R (5'-AGCTTaagcttCTTTGCCGCTTTTTGTTTTCTC-3') were used to amplify the 3' part of Ll.LtrB intron. The lower case letters indicate the appended restriction enzyme sites used for cloning. In the second round of PCR, the two first round PCR products were mixed and amplified by PCR with the outside primers Ex1F and Ex2R. The second round PCR product was digested by EcoRI and HindIII and inserted between the EcoRI and HindIII sites of pBSIISK+ (Stratagene) to yield plasmid pMod4. The *bsd* gene from pCMV/bsd (Invitrogen) was then inserted into this pMod4 to generate pMod4Bsd. For this step, the *bsd* gene was amplified by PCR, using pCMV/bsd as a template, and, BsdEV (5'-ATCGTgatatcTTACATAACTTACGGTAAATGGC-3') and BsdML (5'-

ATGCAacgcgtCAGACATGATAAGATACATTGATG-3') as primers. The PCR product was digested by EcoRV and MluI and inserted into the pMod4 to generate pMod4Bsd.

To generate pACD-Mod4-GFP, the GFP ORF was amplified from pCSGFP2 (provided by Enrique Amaya) using primers pModGFP2s (5'-GCGGCcggaccgATGAGTAAAGGAGAAGAACTTT-3') and pModGFP2as (5'-GCGGCgcttagcTTATTTGTATAGTTCATC-3'). Primer pModGFP2s contains an RsrII restriction site and pModGFP2as contains a BlpI site (the restriction enzyme sites are shown in lower case letters). The PCR product containing the GFP gene was digested with RsrII and BlpI and used to replace the *bsd* gene in pMod4Bsd.

pACD-GFP-RAM and pACD-GFP-RAM1363 contains a wild-type 393-bp *td* intron or mutant *td* intron (*td*1363) inserted within the GFP gene of pACDMod4GFP. The mutant intron contains a G to A mutation at nucleotide residue 67 with the intron's P5 stem element (Mohr et al. 1992). The wild-type *td* intron and mutant *td*1363 intron were amplified from pACD2KanRAM (Zhong et al. 2003) and pTZtd1363 (Belfort et al. 1987), respectively, using primers GFPRAM3s (5'-GTGGTCTCTCTTTTCGTTTGGGTTAATTGAGGCCTGAGTATAAG-3') and GFPRAM2as (5'-CACACAATCTGCCCTTTCGAAAGCATTATGTTCAGATAAGGTCG-3'). The *td* intron was inserted by combining three PCR products. Primer pModGFP2as and GFPRAM3as were used to amplify a segment of pACD-Mod4-GFP upstream of *td* intron insertion site (named fragment A), and primers pModGFP2s and GFPRAM2s were used to amplify a DNA segment of pACDMod4GFP downstream of *td* insertion site (named fragment B). The 3' end of fragment overlaps the 5' end of *td* intron and the 5' end of fragment B overlaps the 3' end of *td* intron. The amplified *td* intron and fragment A and

B were mixed and use as templates for PCR with the outside primers pModGFP2as and pModGFP2s, which contain BlpI and RsrII sites respectively. The combined PCR product was digested with BlpI and RsrII and cloned between the corresponding sites of pACD-Mod4-GFP to generate pACD-GFP-RAM or pACD-GFP-RAM1363.

### 4.9.2 Preparation of Ll.LtrB RNPs

***Transcription***

The Ll.LtrB intron was transcribed from a linearized plasmid or PCR product containing T3 or T7 promoter followed by the Ll.LtrB intron sequence. For lariat introns, the 50 $\mu$g plasmid (pACD3 or pACD-GFP-RAM or derivatives with modified IBS1/2 and EBS1/2 sequences) was linearized by digestion with NheI (New England Biolabs) at 37ºC overnight. The linear plasmid DNA was phenol-chloroform extracted, ethanol precipitated at -20ºC for 2 h and dissolved in 100 µl of distilled $H_2O$. The linear intron RNA was synthesized PCR product generated from pACD3, pACD-GFP-RAM or derivatives with modified IBS1/2 and EBS1/2 sequences using T3-LIS-1G (5'-aattaaccctcactaaaGTGCGCCCAGATAGGGTGTTAAGTCAAG-3') and LtrB940a primers as described previously (Mastroianni et al. 2008). To generate a homogenous 3' end of intron RNA, the LtrB940a primer was purified by HPLC. The transcribed RNA was extracted by phenol-chloroform twice and precipitated with 2.5 M LiCl at -20ºC for 30 min. The RNA was then pelleted at 16,000×g for 15 min at 4ºC, and the pellet was washed twice with 70% ethanol. After air drying, the pellet was dissolved in nuclease-free water and precipitated in the presence of 2.5 M ammonium acetate to further remove the unincorporated nucleotides. The purified RNA was then used for splicing and RNP reconstitution.

***Self-splicing***

The Ll.LtrB intron was self-spliced as previously described (Mastroianni et al. 2008).

To splice the *td* intron in GFP-RAM construct, RNA was spliced at 10-25 mM $Mg^{2+}$ in the presence of 0.2 mM GTP, 20 mM Tris-HCl, pH 7.5, 100 mM KCl at 37ºC for 1 h. Prior the splicing reaction, the RNA was renatured by heating to 55ºC and then slowly cooling to 37ºC in 8 mM $MgCl_2$, 20 mM Tris-HCl, pH 7.5, and 100 mM KCl. The $Mg^{2+}$ and GTP were added to the desired concentration as the temperature reached 37ºC. The RNA was precipitated by 1/10 volume of 3 M sodium acetate and 2 volumes of ethanol at -20ºC for at least 2 h. After ethanol precipitation, the RNA was spun down and washed with 70% ethanol. The RNA pellets were dissolved in 1 ml of nuclease-free water. The splicing reaction was subsequently checked by electrophoresis on a denaturing 4% polyacrylamide gel.

***Preparation of LtrA protein and RNP reconstitution***

The LtrA protein was expressed in *E. coli* BL21 (DE3) using an intein-based affinity purification system (Impact; New England Biolabs) (Saldanha et al. 1999). RNP reconstitution was done with 100 nM self-spliced intron RNA, 200 nM purified LtrA, 40 mM Tris-HCl, pH 7.5, 5 mM $MgCl_2$, and 450 mM NaCl in 10 ml of reaction medium as previously described (Mastroianni et al. 2008). The reaction was incubated at 30ºC for 30 min. After pelleting the RNP particle by ultracentrifugation in a Beckman 50.2 Ti rotor at 145,000 x g for 16 h at 4°C, the pellet was dissolved in 30-100 μl of 40 mM HEPES, pH 7.5, 10 mM KCl and 5 mM $MgCl_2$.

### 4.9.3 Target DNA-primed reverse transcription (TPRT)

TPRT reaction was used to determine the *in vitro* reconstituted RNP activity. The reaction were carried out by incubating 1 μg of unlabeled target plasmid with 1 μl of reconstituted RNP particles in 20 μl of reaction medium [10 mM KCl, 10 mM $MgCl_2$, and 50 mM Tris-HCl (pH 7.5)] containing 0.2 mM dNTP and 20 μCi of [$\alpha$-$^{32}$P]dTTP (3000 Ci/mmol; Perkin-Elmer). The reactions were initiated by addition of the RNP particles, and the mixtures were incubated for 30 min at 37°C. Products were purified and analyzed as previously described (Saldanha et al. 1999).

### 4.9.4 Nucleoplasmin purification

*E. coli* BL21(DE3) was transformed with the plasmid expressing CORE8D (Banuelos et al. 2003) and plated on LB medium containing ampicillin. For starter cultures, a single colony from the plates was inoculated into 50 ml of LB medium containing ampicillin and incubated overnight at 37ºC. 10 ml of the overnight culture were then inoculated into 1 liter of LB containing ampicillin in a 4-liter flask and shaken vigorously at 37ºC, until the cell density reached $O.D._{595}$ 0.6. The cultures were then shifted to 18ºC and induced with 1 mM IPTG overnight. The cultures were harvested by centrifugation in a Beckman JA-14 rotor (4,000×rpm for 10 min), and the resulting bacterial pellet was resuspended in a homogenization buffer (25 mM Tris-HCl pH 7.5, 120 mM NaCl, 1mM DTT) supplemented with phosphatase inhibitors (10 mM beta-glycerophosphate, 10 mM Na molybdate, 100 mM Na vanadate), protease inhibitors (0.2 mM AEBSF and one Roche Protease inhibitor cocktail tablet without EDTA) and 0.1 mg/mL lysozyme. 15ml of this homogenization buffer was used to resuspend each liter of cell pellet. Cells could then be stored at -70ºC for months. Cells to be lysed were thawed and incubated on ice for 1 h and then subjected to two or three cycles of freezing and

thawing between -70ºC and 25ºC, followed by sonication (Branson 450 Sonifier, Branson Ultrasonics Inc., Danbury, CT; four 10 s bursts at amplitude 60, with 10 s between bursts). After sonication, insoluble material was pelleted by centrifugation in a Beckman JA20.1 rotor (11,000×rpm for 15 min). The supernatant was further heated at 70ºC for 15 min with slight shaking from time to time, and insoluble material was pelleted by centrifugation at 100,000×g for 1 h. The clear supernatant was applied directly to a Q-Sepharose column in 25 mM Tris-HCl, pH 7.5, 0.2 mM PMSF (phenylmethanesulphonylfluoride), and proteins were eluted with a gradient of 100 mM to 1 M NaCl in the same buffer. The eluted protein was further purified by gel-filtration though a Superdex 200 16/60 column in 25 mM Tris-HCl, pH 7.5, 100 mM NaCl.

### 4.9.5 Group II intron (Ll.LtrB)-mediated transgenesis by sperm nuclear transplantation into unfertilized eggs

Preparation of buffers, transplantation apparatus, needles, agarose-coated injection dishes and frogs was as described previously (Kroll and Amaya 1996). The reaction was prepared by adding 4 $\mu$l sperm nuclei stock (1~2×$10^5$ nuclei) to a reaction mix consisting of 3.6 µl of 5× Sperm Dilution Buffer (SDB; 250 mM sucrose, 75 mM KCl, 0.5 mM spermidine trihydrochloride, 0.2 mM spermine tetrahydrochloride, adjusted to pH 7.3-7.5 with NaOH), 3 µl of 2 mM dNTP, 2 µl of 150 mM $MgCl_2$, indicated amounts of purified nucleoplasmin (Core8D NP) and RNP. The Core8D NP concentration was optimized for different target sites. The total reaction volume was adjusted to 27 $\mu$l with distilled water. The reaction was mixed gently by pipetting up and down with a wide bore tip, and the mixture was incubated for certain amount of time at temperatures specified for individual experiments. After the incubation, 5 µl of reaction mix was diluted into 150 µl SDB with 5 mM $MgCl_2$ at room temperature. The treated sperm nuclei were injected into eggs as described previously (Kroll and Amaya 1996).

Fertilized embryos were kept in an 18ºC incubator. When healthy embryos reached the 4-cell stage (Nieuwkoop 1994), they were transferred into 60×15 mm petri dishes containing 0.3×Marc's Modified Ringers (MMR) + 4% Ficoll + 50 $\mu$l/ml gentamycin. The culture medium was changed 3~4 times per day, and dead embryos were periodically removed from the plates. When the embryos reached stage 12, the medium was replaced with 0.3×MMR + 50 $\mu$g/ml gentamycin without Ficoll. The embryos were kept in this media until they hatched with 2 or 3 buffer changes per day. After hatching, the embryos were transferred into a 10 cm × 20 cm plastic container with 0.3× MMR at room temperature. Tadpoles were fed with Nasco Frog Brittle for tadpole *Xenopus* (Nasco) every other day. Water was changed 3~4 h after every feeding.

### 4.9.6 Tail clipping from tadpoles

Tadpoles were anaesthetized by transferring them to dishes containing 0.025% ethyl 3-aminobenzoate methanesulfonate salt (Tricaine; Sigma) in 0.3× MMR. Anaesthesia usually took 1 to 2 min, depending on the size of the tadpole. A sterile scalpel was then used to cut away the posterior 1/4 to 1/3 of the tail (as measured from the posterior of the gut to the tip of the tail). A clean forceps was used to transfer the tail sample out of the dish. Using a clean teaspoon, tadpoles were returned to tanks containing 1 to 2 inches of 0.3× MMR to recover. Recovery usually took 5-10 min, after which the tadpoles were able to swim and feed normally. Regeneration of the tail took around five days. The DNA from tail was extracted by using a Mammalian Genomic DNA Extraction kit (Sigma).

### 4.9.7 PCR analysis of intron integrations into target sites

Ll.LtrB intron integration efficiency in sperm TPRT reactions was analyzed by real-time PCR. After the sperm TPRT reaction, RNPs were inactivated by heating at 75ºC

for 10 min, and the reaction mixture was then incubated with 0.5 μl of RNase H (2 units/μl; Invitrogen) at 37°C for 30 min to remove the RNA strand of RNA/cDNA hybrid nucleic acid. The sperm DNA was then extracted by using a Mammalian Genomic DNA Extraction kit (Sigma).

Two sets of primers, one for targeted sites and the other for untargeted sites, were used to determine intron integration efficiency. For targeted sites, a sense strand primer annealing 4-bp upstream of the intron's 3' end and an antisense gene-specific primer complementary to a sequence 120-bp downstream of the intron insertion site were used to quantitate the amount of intron-target site 3' junction in reacted sperm DNA. For untargeted sites, a set of primers amplifying a segment of the mitf exon 6 was used. Because mitf is single copy gene in *Xenopus*, the copy number of mitf gene is equal to the number of sperm nuclei in the reaction. The primers were designed with similar GC content and $T_m$ value to minimize differences in amplification efficiency between primer sets. The intron integration efficiency was calculated using the equation specified for each experiment (See Figure 4.4D and 4.12C).

To check 3' intron integration junctions in sperm TPRT or embryos, an intron-specific primer LtrB816s (5'-CAGTGAATTTTTACGAACGAACAATAAC-3'), which anneals to a sequence 120-bp upstream of intron 3' end, and a gene specific primer complementary to a sequence located 150~300 bp downstream of intron insertion site were used. If nested PCR was required, a nested intron specific primer was LtrB890s (5'-GTGCAAACCAGTCACAGTAATGTG-3') and a nested gene specific primer anneals a few bases "internal" to the first primer binding sites.

To detect the 5'-integration junctions in transgenic embryos, a gene specific primer located upstream of intron-insertion site, and an intron specific anti-sense primer LtrB256a (5'-CATTTTGAGGTTTTCCTCCCTAATC-3') were used.

To check the intron integration in the plasmid target site in *Xenopus* oocyte, both 5'- and 3'-integration junctions were checked by PCR. For 5' integration junction, the primers were ForpBRR (5'-CTGATCGATAGCTGAAACGC-3'), which anneals to a sequence upstream of the intron target site, and an antisense primer LtrB174a (5'-CGGCTCTGTTATTGTTCGTTCG-3'), which anneals to 112 bp upstream of the intron 3' end. The 5'-integration junction PCR product was sequenced using the primer Rseq (5'-CCATGCGAGAGTAGGGAAC-3'). For 3'-integration junction, the primers were LtrB816s, which anneals to a sequence 120-bp upstream of intron's 3' end, and RevpBRR (5'-AATGGACGATATCCCGCA-3'), which anneals to a sequence downstream of the intron target site in the recipient plasmid. The 3' junction product was sequenced using the primer LtrB816s.

### 4.9.8 TAIL PCR

TAIL PCR used to identify intron integration sites in the *Xenopus* genome was as previously described (Liu et al. 1995). The 5'-integration junctions were amplified using two intron specific primers, LtrBtailP1 (5'-CCCTTGCAAGATTTTCAAGCTCTAGTGC-3') for the first PCR step and LtrBhitailP2 (5'-ACGATGGACTCCAGTCCGGCCATTAGGCATTCTTGTTTAGGGTATCCCCAG-3') for the second PCR step. The 3' integration junction were amplified using two intron specific primers, HiTailLtrB816s (5'-GTGAATTTTTACGAACGAACAATAACAGAG-3') in the first PCR step and

HiTailLtrB890s (5'-ACGATGGACTCCAGTCCGGCCGTGGTGCAAACCAGTCACAGTAATGTG-3') for the second PCR step. The degenerate primers used in the reaction are LAD1 and LAD2, as described by Liu et al. 1995. The final PCR product was gel-purified and sequenced using the same intron-specific primer used for the second PCR step.

### 4.9.9 DNA endonuclease assay

DNA endonuclease assays were carried out with a 129-bp DNA substrate, which was internally labeled by PCR in the presence of [$\alpha$-$^{32}$P]dTTP (3,000 Ci/mmol; Perkin-Elmer, Waltham, MA) (San Filippo and Lambowitz 2002), or with 60-bp DNA oligonucleotide substrates labeled at the 5' end of the top or bottom strand with [$\gamma$-$^{32}$P] ATP and phage T4 polynucleotide kinase (New England Biolabs). The 60-bp DNA substrates were prepared by annealing complementary synthetic oligonucleotides and purified using a Centri-Sep spin column (Princeton Separations, Adelphia, NJ). For assays with the internally labeled DNA substrate, the DNA (1.5 nM; 56,000 cpm) was incubated with RNPs (815 nM) for 30 min at 37ºC in 20 μl of reaction medium containing 10 mM KCl, 10 mM $MgCl_2$, and 50 mM Tris-HCl, pH 7.5. For assays with the 5'-labeled DNA substrates, the DNA (50 nM) was incubated with RNPs (100 nM) for 30 min at 37ºC in 20 μl of the same reaction medium plus 0.2 mM each of dATP, dCTP, dGTP, and dTTP (referred to as dNTPs). After the incubations, the reactions were terminated by extraction with phenol-chloroform-isoamyl alcohol (25:24:1; phenol-CIA) followed by ethanol precipitation, and the products were analyzed by electrophoresis in denaturing 6% or 10% (w/v) polyacrylamide gels, which were dried and scanned with a phosphorimager.

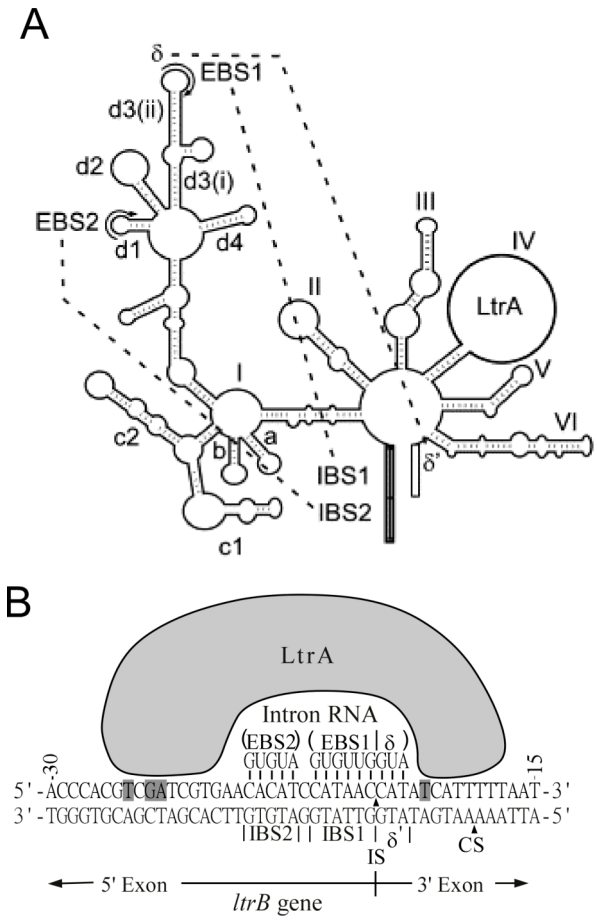



Figure 4.1: The Ll.LtrB group II intron and its DNA target site recognition mechanism.

(A) Predicted secondary structure of the Ll.LtrB intron RNA. The intron RNA consists of six double-helical domains (DI-VI) radiating from a central wheel. The EBS1/IBS1, EBS2/IBS2, and δ-δ' base-pairing interactions between the intron and flanking exons are indicated by dashed lines. The IEP (LtrA protein) is encoded in DIV. (B) DNA target site recognition by Ll.LtrB RNPs. The EBS1, EBS2, and δ sequences in DI base pair with complementary sequences IBS1 and IBS2 in the 5'-exon and δ' in the 3'-exon. The LtrA protein recognizes specific nucleotide residues in the distal 5'-exon and 3'-exon regions, including T-23, G-21, A-20, and T+5 (highlighted with gray background). IS and CS denote the intron insertion site and the LtrA cleavage site, respectively.

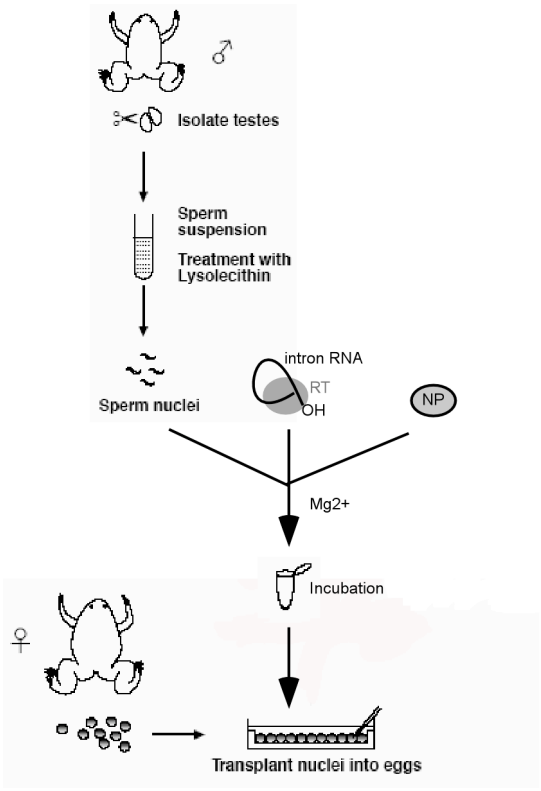



Figure 4.2: Group II intron-based transgenesis in *X. laevis*.

Sperms are isolated from male frog testes and are treated with lysolecithin to permealize the sperm nuclei. The sperm nuclei are then incubated with *in vitro* reconstituted group II intron RNP particles and nucleoplasmin (NP) in the presence of magnesium ($Mg^{2+}$) to achieve site-specific intron integration into the sperm genome. The reaction mixture is diluted to optimal sperm nuclei concentration for *in vitro* fertilization and injected into freshly squeezed unfertilized eggs to generate transgenic frogs.

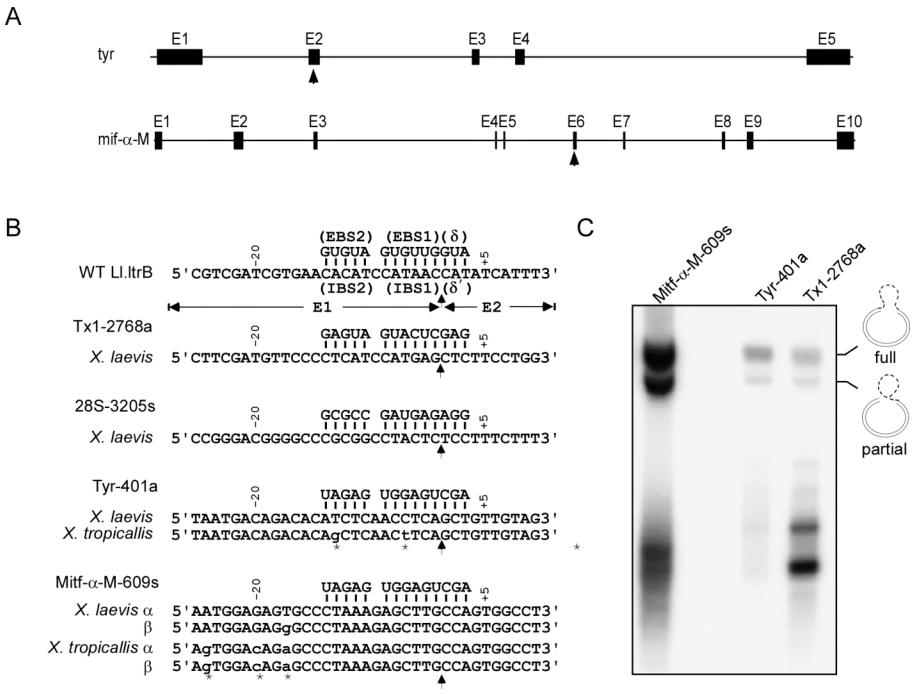



Figure 4.3: Ll.LtrB intron target sites in the *X. laevis* genome and RNP activities in *in vitro* TPRT assays.

(A) Diagram of *X. laevis* tyr and mitf genes. Retargeted introns were designed to insert into exon 2 of the tyr gene and exon 5 of the mitf gene. The black boxes represent the exons and the black lines denote the introns of these genes. The intron-insertion sites in the genes are indicated by arrows. (B) DNA target site sequences in the *X. leavis* genome. The base pairing between the EBS1, 2 and δ sequences of the retargeted introns (targetrons) and their DNA target sites are shown. The "a" or "s" in the targetron names indicates that the intron inserts into the antisense or sense strand of the gene, respectively. The arrows represent the intron-insertion site. Introns are designed based on *X. laevis* gene sequences and the corresponding regions in *X. tropicalis* are also listed. Nucleotide sequences that differ between *X. laevis* and *X. tropicalis* are shown in lower case letters with asterisks below. (C) *In vitro* TPRT assay of reconstituted targetron RNPs. RNPs were incubated with unlabeled target plasmids in the presence of [α-$^{32}$P]dTTP, other dNTPs and TPRT buffer to support reverse transcription (Saldanha et al. 1999). After incubating at 37ºC for 30 min, the reaction was stopped by phenol-chloroform extraction. Unincorporated [α-$^{32}$P] dTTP was removed by running the extracted reaction mixture through a Micro Bio-Spin 30 Column (Biorad). The eluted products were analyzed in a 1% agarose gel, which was dried and scanned with a phosphorimager. Schematics of partial and full reverse splicing products are shown on the right of the gel. DNA and intron RNA are shown as solid double lines and dashed lines, respectively. The smaller size products in the gel may be due to cDNA synthesis on RNA fragments in the preparation.

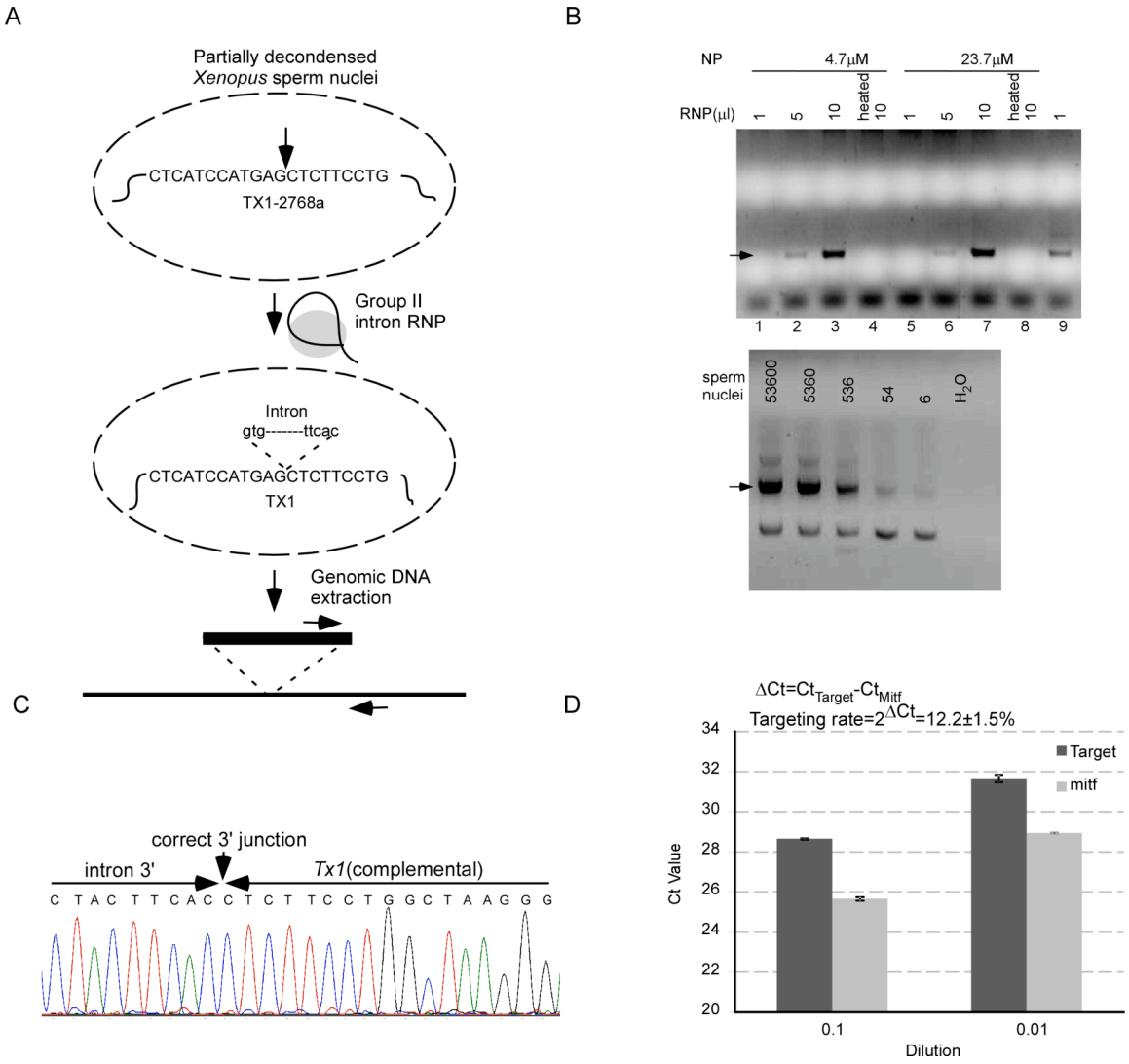



Figure 4.4: Tx1-2768a intron targeting in *X. laevis* sperm nuclei.

(A) Schematic of intron targeting in *X. laevis* sperm nuclei. Sperm nuclei were partially decondensed with purified nucleoplasmin, and the decondensed sperm nuclei were incubated with *in vitro* reconstituted group II intron RNPs. After terminating the targeting reaction by heat inactivation, genomic DNA was extracted for PCR detection of the 3'-integration junction. (B) PCR detection of the Tx1-2768a intron integration junction in sperm nuclei. Different amounts of RNPs and nucleoplasmin were used in sperm TPRT reactions to achieve the highest targeting efficiency. The intron and target site junction can be detected in the reactions with 5 or 10 μl of RNP (1.8 mg/ml lariat RNA based on $O.D._{260}$). However the targeted product can't be detected in the sperm reaction with 1 μl of RNP. As the positive control, the RNP was incubated with extracted

sperm genomic DNA (lane 9). Parallel reactions with heat-inactivated RNPs did not show the junction (lane 4 and 8). The insertion efficiency is proportional to the amount of RNP used in the reaction. At the bottom, the intron-integration efficiency was determined by PCR of the serially diluted sperm reaction DNA. The amount of sperm nuclei in the PCR reaction was determined as described in Methods. The identity of correct-sized PCR product (arrow) was confirmed by DNA sequencing. Site-specific intron integration could be detected in as few as 54 sperm nuclei for the Tx1-2768a intron. All other bands detected were also sequenced and found to be PCR artifacts (target site without integration or unknown sequences). (C) Sequence analysis of the 3'-integration junction by Tx1-2768a RNP. (D) Quantitative analysis of Tx1-2768a intron targeting in sperm nuclei by real-time PCR. The DNA from a sperm TPRT reaction was extracted by a Mammalian Genomic DNA Extraction kit (Sigma) and eluted in 100 μl $H_2O$. The amount of DNA used in the real-time PCR is 3 μl of diluted DNA from the elution. The Ct values are plotted as a bar graph for dilutions that give optimal PCR amplification efficiency. PCR product "Target" detects the 3'-intron integration junction. PCR product "Mitf" detects the single copy gene mitf, which is equivalent to the number of sperm nuclei in the reaction. The bar graphs show the mean for three determinations. The error bar indicates the standard deviation. The equation used to calculate the intron targeting efficiency is shown in graph.

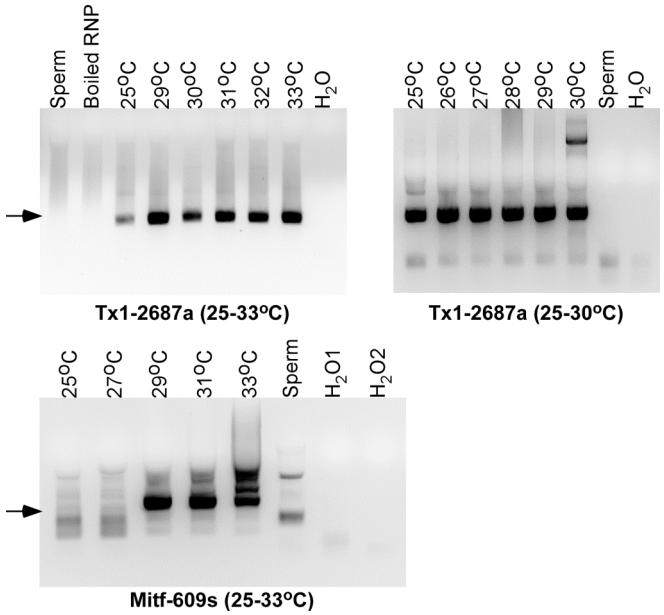



Figure 4.5: Temperature optimization of sperm TPRT reaction for targetron Tx1-2768a and Mitf-α-M-609s.

Tx1-2768a or Mitf-α-M-609s targetron RNPs were incubated with sperm nuclei, nucleoplasmin, dNTP and $Mg^{2+}$ at temperatures ranging from 25ºC to 33ºC. Reactions were stopped by heat inactivation and DNA was extracted for PCR analysis. To detect the targeted product, an intron-specific primer and a target site-specific primer were used to amplify the specific 3'-integration junction. The arrows indicate the PCR product corresponding to the correct 3'-integration junctions, whose identities were confirmed by sequencing. For Tx1-2768a, the targeted product can be detected at all temperatures ranging from 25ºC to 33ºC. For Mitf-α-M-609s intron, the targeted product can be detected only at 29ºC or higher by nested PCR. Genomic DNA from untreated sperm nuclei denoted "Sperm" and PCR blank controls denoted "$H_2O$" were negative controls for the PCR reactions. The "boiled RNP" represents the sperm TPRT reaction using the heat inactivated RNPs. Other bands on the gels were PCR artifacts or due to primer dimers.

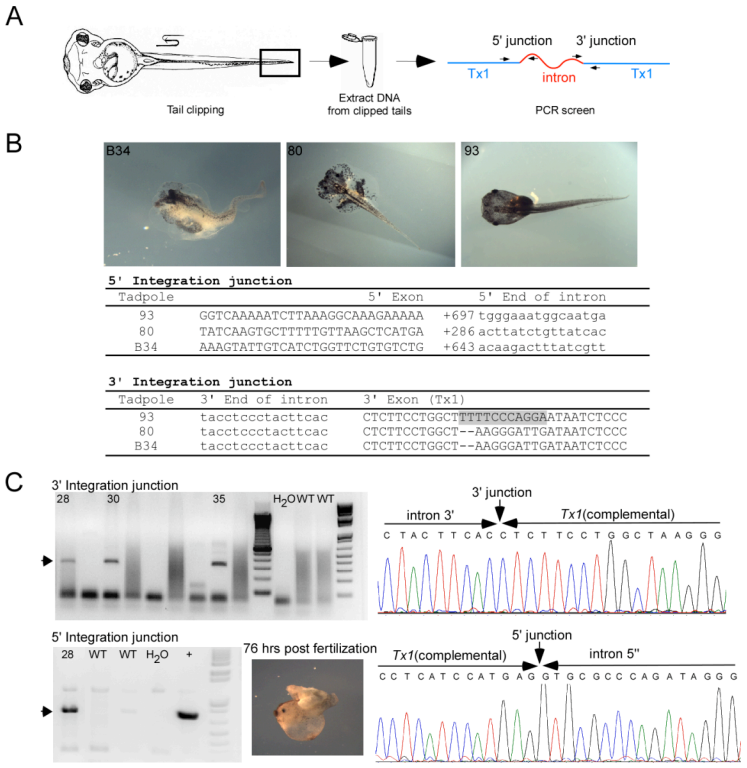



Figure 4.6: Transgenic tadpoles with Tx1-2768a intron integrated in the Tx1 gene.

(A) Schematic of procedure used to screen for tadpoles with a targeted group II intron inserted into the Tx1 gene. The genomic DNA of an individual tadpole was extracted from the clipped tail, which was removed by a tail-clipping surgery. PCR was used to detect tadpoles with both 5'- and 3'- intron integration junctions. Primers used for PCR detection are indicated as arrows. The Tx1 gene and inserted intron are shown as blue and red lines, respectively. (B) Tx1 transgenic tadpoles with a targeted group II intron inserted into the Tx1 gene. 2,200 eggs were injected with sperm nuclei that had been incubated with the Tx1-2768a targetron. Forty eight of them developed to the tadpole stage. Tadpoles B34, 80 and 93 had intron integrations at the correct site in the Tx1 gene. The sequences at the 5' and 3' integration junction are shown below. Small sequence differences (highlighted in gray color) near the 3'-junctions in the three targeted tadpoles may reflect polymorphism of different Tx1 elements and/or misincorporation errors during a DNA repair process used for cDNA integration. The sequences in introns and exons were in upper and lower cases respectively. The numbers ahead the integrated intron 5' ends were the nucleotide positions in the full length intron. (C) Tx1 transgenic tadpoles with full length Tx1-2768a intron integrated into the target site. In this experiment 1,500 eggs were injected. Three embryos were found to have the correct 3'-integration junction. Among them, embryo 28 was also found to have the correct 5'-integration junction. The picture shows embryo 28 at 76 h post fertilization. Sequencing results for the 3'-and 5'-integration junctions are shown on the right.

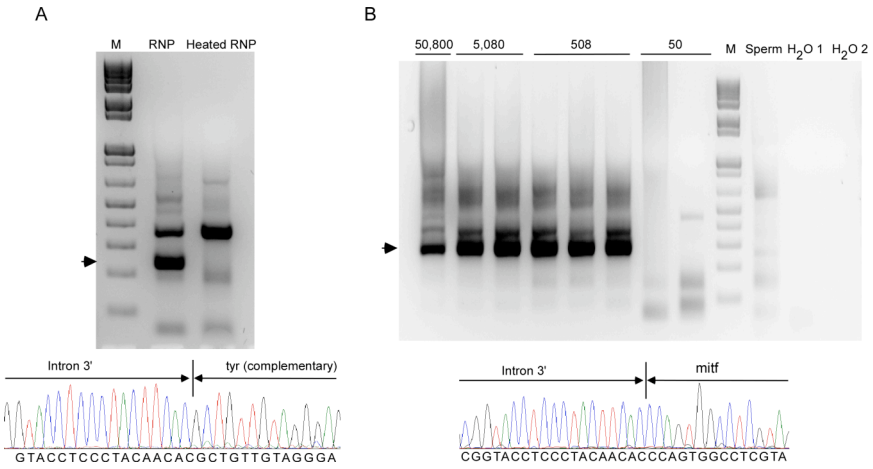



Figure 4.7: Targeting the tyr and mitf gene in *X. laevis* sperm nuclei.

*In vitro* sperm TPRT reactions and PCR showing the site-specific integration of retargeted Ll.LtrB-ΔORF introns into single copy protein coding genes, tyr and mitf. (A) Sperm TPRT of Tyr-401a RNP. Sperm nuclei were incubated with purified nucleoplasmin and Tyr-401a RNP. A parallel reaction with heat-inactivated RNPs was used as a negative control. After incubation for 30 min at 33ºC, total nucleic acid was extracted and analyzed by nested PCR to detect the 3' intron integration junction. The correct-sized PCR product (arrow) was detected in cells incubated with active RNPs, and its identity was confirmed by DNA sequencing. (B) Sperm TPRT of Mitf-α-M-609s RNPs. The Mitf-α-M-609s RNPs were incubated with sperm nuclei and nucleoplasmin at 30ºC for 15 min. After the reaction, nucleic acids were extracted and diluted for nested PCR. The number of sperm nuclei for each sample was determined, as described in Methods. The PCR product corresponding to the correct 3' integration junction is indicated by the arrow, and its identity was confirmed by sequencing (below). Lane "Sperm" refers to the sperm TPRT reaction without Mitf-α-M-609s RNPs. Lane "$H_2O$ 1" and "$H_2O$ 2" are the distilled water controls for the first and second step of the nested PCR. All other bands detected were also sequenced and found to be PCR artifacts (target site without an integrated intron or unknown sequences). M: 1 Kb Plus DNA Ladder (Invitrogen).

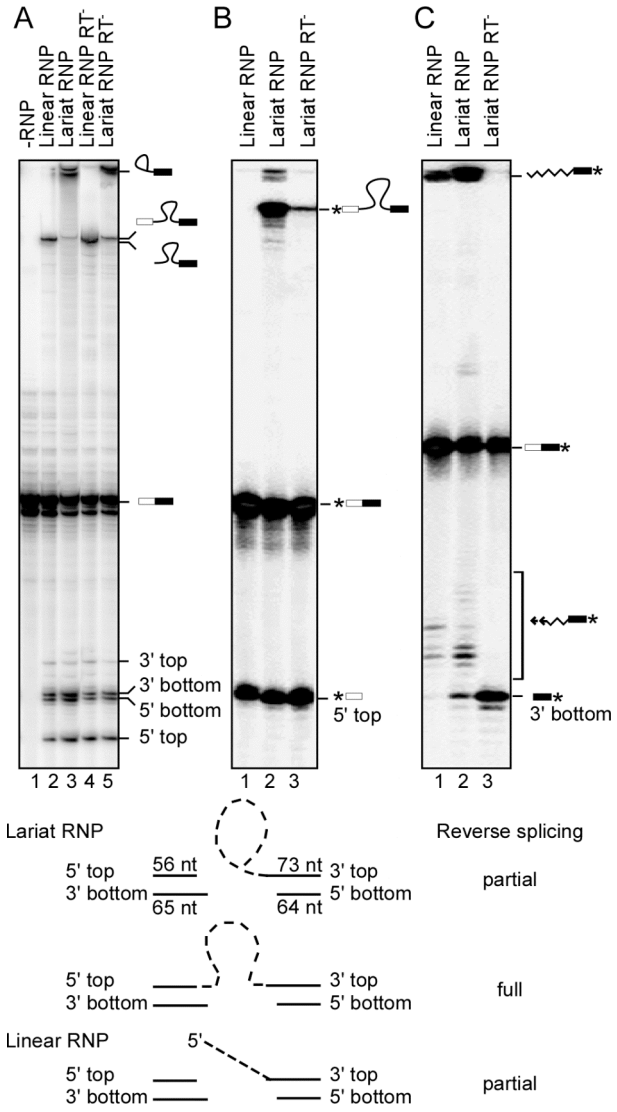



Figure 4.8: Target DNA cleavage and target DNA-primed reverse transcription reactions of RNPs reconstituted with lariat or linear Ll.LtrB intron RNA.

Ll.LtrB RNPs containing lariat or linear intron RNA were incubated with $^{32}$P-labeled DNA oligonucleotide substrates containing the Ll.LtrB target site (positions -56 to +73 from the Ll.LtrB intron-insertion site), and the products were analyzed in a denaturing polyacrylamide gel. (A) Lariat and linear RNPs were incubated with internally labeled DNA substrate for 30 min at 37ºC, as described in Methods. (B) and (C) Lariat and linear RNPs incubated with DNA substrates labeled (asterisk) at the 5'-end of the top and bottom strand, respectively in the presence of 0.2 mM dATP, dCTP, dGTP, and dTTP for 30 min at 37ºC. In (A)-(C), products are indicated to the right of the gel. The schematic at the bottom diagrams the products expected for each reaction. All lanes are from the same gel, but some lanes in (A) were rearranged to be adjacent.

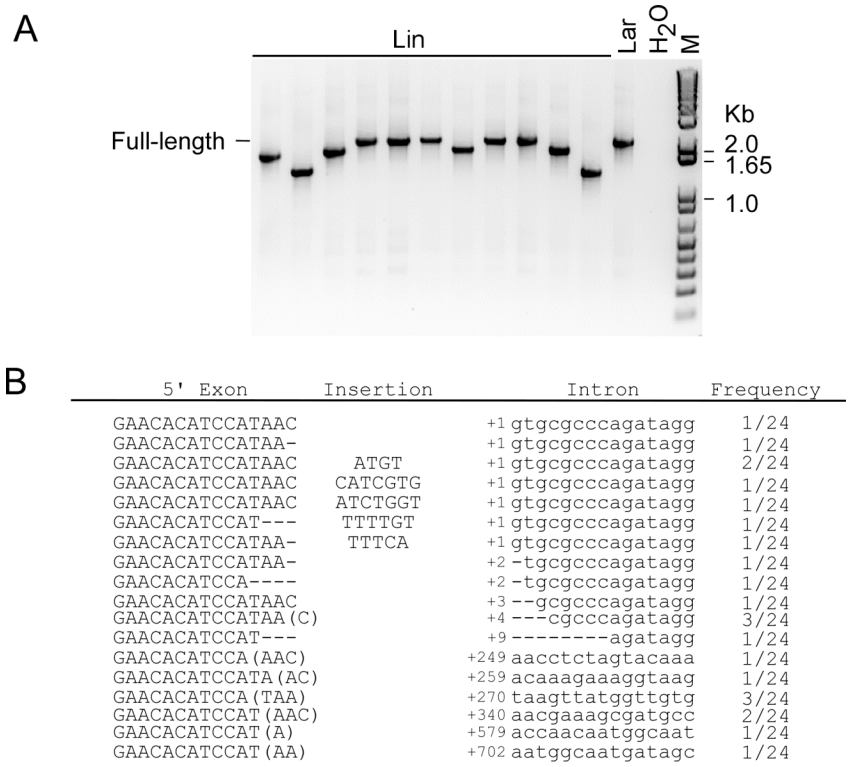
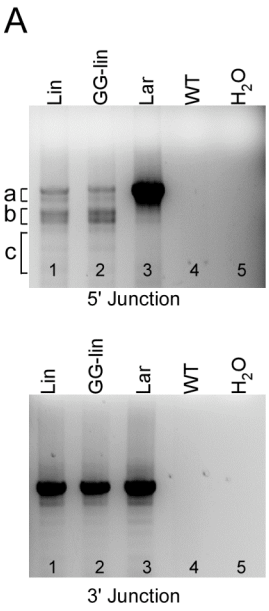



Figure 4.9: Retrohoming of linear Ll.LtrB intron RNPs in *X. laevis* oocyte nuclei analyzed using an *E. coli* genetic assay for retrohoming products.

A recipient plasmid containing the Ll.LtrB target site cloned upstream of a promoterless $tet^R$ gene and RNPs containing a linear Ll.LtrB-ΔORF intron RNA with a phage T7 promoter near its 3' end were injected sequentially into *X. laevis* oocyte nuclei. After incubation for 2 h at 25ºC, nucleic acid was extracted and transformed into *E. coli* HMS174(DE3), which was then plated on LB agar containing ampicillin with or without tetracycline. Site-specific integration of the intron containing the T7 promoter into the target site activates the $tet^R$ gene, and integration efficiency is calculated as the ratio of ($Tet^R$ + $Amp^R$/$Amp^R$ colonies). The integration efficiency for linear RNPs in this experiment was 0.04% compared to 3% for lariat RNPs. Site-specific integration was confirmed by the colony PCR using primers For-pBRR and Rev-pBRR flanking the target site in the recipient plasmid, and the 5' integration junction was sequenced by using the primer Rseq. (A) Colony PCR of eleven $tet^R$ colonies using primers For-pBRR and Rev-pBRR flanking the recipient plasmid target site. A PCR product of 1.9 kb is expected for integration of the full-length intron. Smaller PCR products reflect integration of 5'-truncated introns. Colony PCR of a $tet^R$ colony obtained with lariat RNPs (LA) is shown for comparison. Lane M, 1-kb plus DNA ladder (Invitrogen). (B) Sequencing of 5'-integration junctions. Six of sixteen colonies showed precise insertion

of the full-length intron; two colonies had mutations at the 5' junction, presumably reflecting inaccurate DNA repair; five had 5' truncated introns; two had 12-nt insertion at the 5' junction. The 5' exon and intron sequences are shown in upper and low case respectively. The number before the first intron nucleotide indicates its nucleotide position in the full-length intron. Mutant nucleotides in the intron are highlighted in gray. Homologous sequences in the 5' exon, which are not present in the retrohoming product, are shown in parentheses. The experiment was done in collaboration with Marta Mastroianni.




B

| | 5' Exon | Insertion | Intron | Frequency |
|---|---|---|---|---|
| Band a | GAACACATCCAT--- | TCAC | +1 gtgcgcccaggtagg | 1/25 |
| | GAACACATCCATAAC | C | +1 gtgcgcccagatagg | 1/25 |
| | GAACACATCCATAAC | CCTAATACACT | +1 gtgcgcccagatagg | 1/25 |
| | GAACACATCCATAAC | | +3 --gcgcccagatagg | 2/25 |
| | GAACACATC------ | | +3 --gcgcccagatagg | 1/25 |
| | GAACACATCC----- | | +5 ----gcccagatagg | 1/25 |
| | GAACACATCCATAA(C) | | +6 -----cccagatagg | 1/25 |
| | GAACACATCCATAA- | TTATATA | +7 ------ccagatagg | 1/25 |
| | GAACACATCCA---- | | +10 gatagggtgttaagt | 1/25 |
| | GAACACATCCA(TAAC) | | +54 taacacagaaaacag | 1/25 |
| Band b | GAACACATCCAT--- | | +244 tctgaaacctctagt | 1/25 |
| | GAACACATCCAT(AAC) | | +249 aacctctagtacaaa | 1/25 |
| | GAACACATCCAT(AA) | | +266 aaggtaagttatggt | 1/25 |
| | GGTGGCGAATATGAA-43 | TTGTGT | +257 gtacaaagaaaggta | 1/25 |
| | GAACACATCCATA(AC) | | +300 accacatttgtacaa | 1/25 |
| | GAACACATCC----- | CA | +303 acatttgtacaatct | 1/25 |
| | GAACACATCC(AT)- | | +305 atttgtacaatctgt | 1/25 |
| | GAACACATCCATAAC | | +359 atctgaatttaccaa | 1/25 |
| | GAACACATCCATAAC | TAAAATTACCCC | +482 gtagtctgagaaggg | 1/25 |
| | AAAGGCTCAGTCGA(A)-285 | | +9 agatagggtgttaag | 1/25 |
| Band c | GAACACATCCAT(AAC) | | +578 aaccaacaatggcaa | 1/25 |
| | GAACACATCCAT--- | GACATCCA | +665 tacgtccagatattt | 1/25 |
| | GAACACATCCATAAC | AT | +699 ggaaatggcaatgat | 1/25 |
| | GAACACATCCATAAC | AC | +702 aatggcaatgatagc | 1/25 |

Figure 4.10: Retrohoming of linear Ll.LtrB intron RNPs in *X. laevis* oocyte nuclei analyzed by direct PCR analysis of integration junctions.

*In vitro* reconstituted intron RNPs and recipient plasmids containing intron target sites were injected into the *X. laevis* oocyte nuclei, as described in Methods. After incubating the injected oocytes at 25ºC for 2 hrs, nucleic acids were extracted. (A) PCR analysis of 5'- and 3'-integration junction. For the 5'-integration junction, For-pBRR, annealing to the upstream of the intron target site, and LtrB174a, an intron specific antisense primer, were used. For the 3'-integration junction, an intron specific sense primer (LtrB816s) and an antisense primer (Rev-pBRR) annealing to the downstream of the intron target site were used. Lanes: (1) Linear RNP; (2) GG-Lin RNPs; (3) Lariat RNP; (4) Wild-type oocyte genomic DNA; (5) distilled $H_2O$. (B) Sequence analysis of 5' junction of linear intron retrohoming products. For the 5'-integration junction, the PCR for retrohoming products of linear RNP resulted in two major bands (bands "a" and "b"). Gel-purified DNA was cloned into Topo TA cloning vector, pCRII-topo (Invitrogen) and the inserted DNA fragments were sequenced. The 5'-exon and intron sequences are shown in upper and low case respectively. The position of first intron nucleotide is indicated at the start of the intron sequence. Mutations in the intron are highlighted in gray. Homologous sequences in the 5' exon, which are not present in the retrohoming product, are shown in parentheses. The experiment was done in collaboration with Marta Mastroianni.

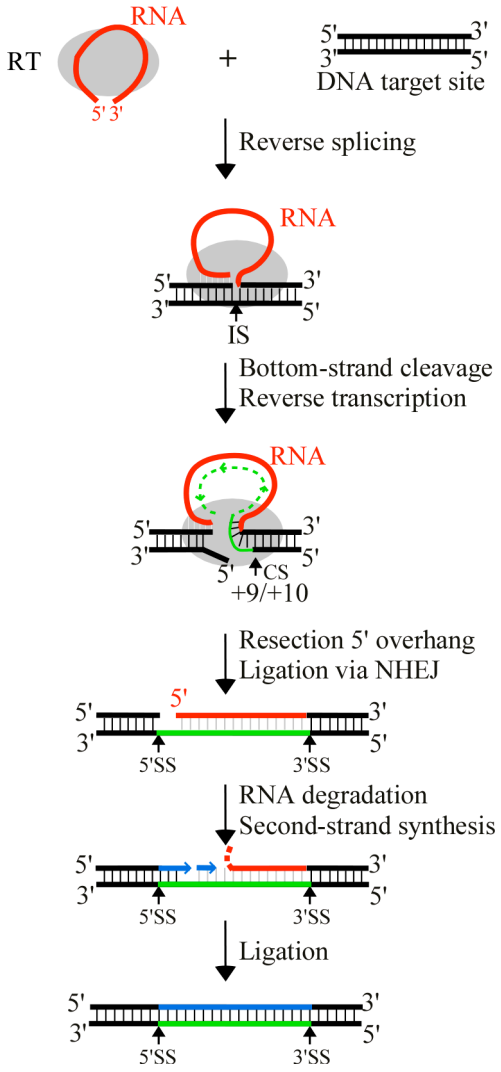



Figure 4.11: Model for retrohoming of linear group II intron RNA.

RNPs containing linear group II intron RNA recognize DNA target sites and carry out the first step of reverse splicing into the intron-insertion site (IS), resulting in ligation of the 3' end of the intron RNA to the 5' end of the 3' exon DNA. The IEP then uses its En domain to cleave the bottom strand between positions +9 and +10 of the 3' exon (CS) and reverse transcribes the attached linear intron RNA. After resection of the 5' overhang resulting from the initial staggered double-strand break at the target site, the intron cDNA is ligated to the 5' exon via non-homologous end joining (NHEJ). The ligation step is inaccurate and often occurs with loss of 5'-exon sequences due to excessive resection, insertion of 5'-truncated introns due to incomplete cDNA synthesis, and/or insertion of

extra nucleotide residues at the ligation junction, presumably due to DNA synthesis by a repair DNA polymerase. Retrohoming is completed as for lariat RNA by degradation or displacement of the intron RNA template strand, second-strand DNA synthesis, and sealing of nicks using host enzymes. Intron RNA, red; LtrA, gray; recipient DNA, black; cDNA, green; second-strand DNA, blue.

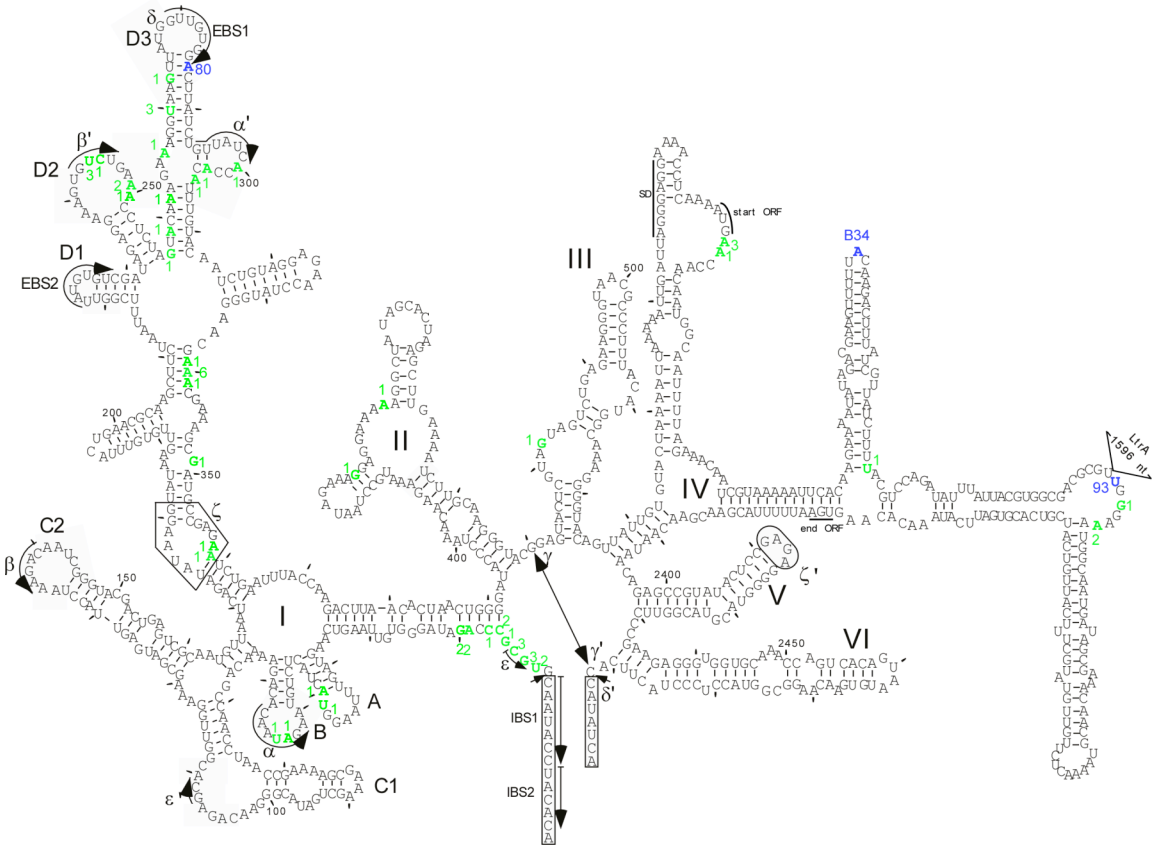



Figure 4.12: The 5' ends of truncated Ll.LtrB introns integrated into the plasmid target site in *X. laevis* oocyte injection and *X. laevis* transgenic tadpoles.

The 5' ends of truncated introns are indicated on the predicted secondary structure of Ll.LtrB RNA. The highlighted green and blue nucleotides are the 5' ends of truncated introns in *X. laevis* oocyte injection experiments and Tx1-2768a transgenic tadpoles generated in Figure 4.6B respectively. The green numbers near the green nucleotides represent the number of times the same truncation occurred in the *X. laevis* oocyte injection. The tadpole names (blue) were also shown near the 5' end of its cDNA. The intron structure consists of six double-helical domains (DI-DVI). The 5'- and 3'-splice sites are indicated by arrows. The Shine-Dalgarno (SD) sequence, initiation and termination codons of the intron ORF, and sequences involved in tertiary interactions (Greek letters) are underlined along the sequences.

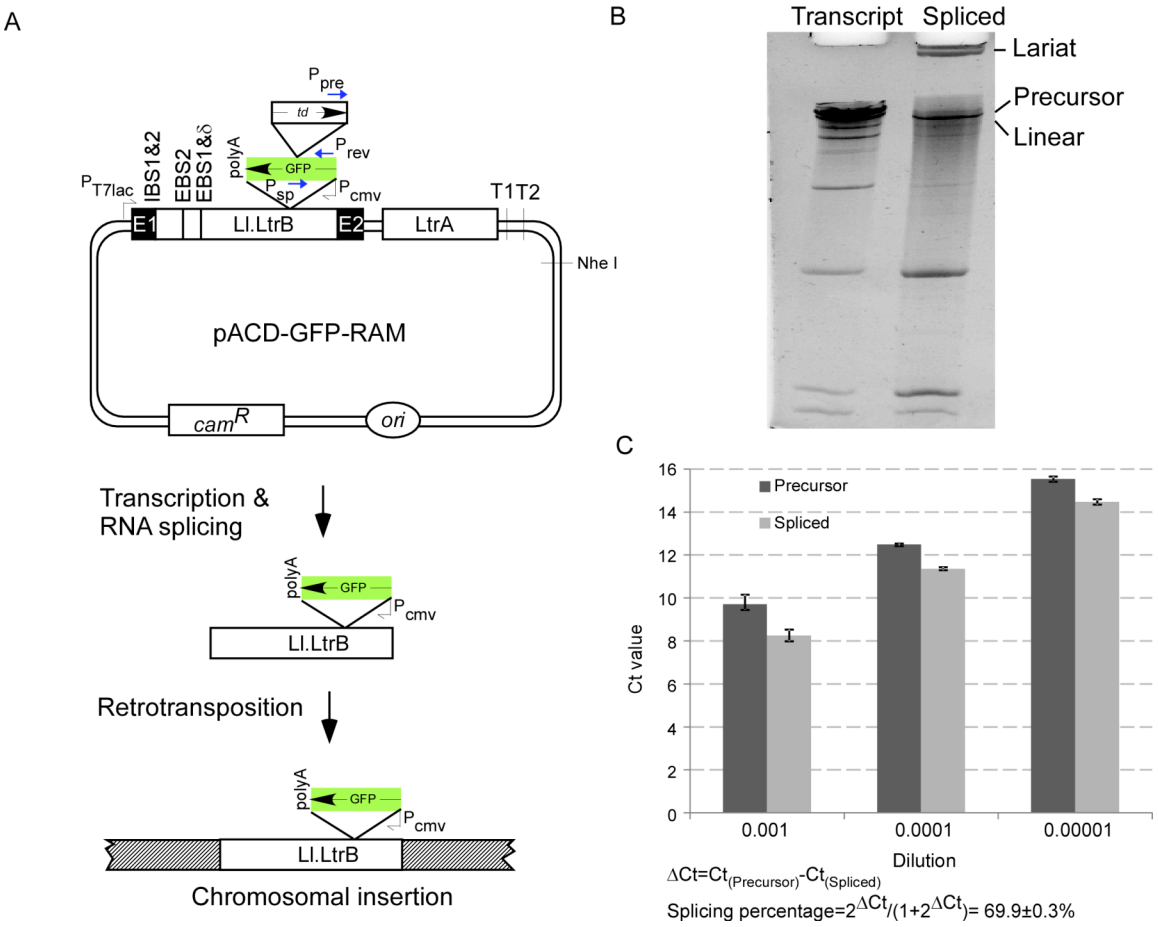



Figure 4.13: The Ll.LtrB-GFP-RAM targetron construct and its splicing activity.

(A) pACD-GFP-RAM constructs. pACD-GFP-RAM is a derivative of pACD3, which contains a 0.9 kb Ll.LtrB-ΔORF group II intron and flanking exons, with the intron-encoded protein, LtrA, expressed from a position downstream of the 3' exon (Guo et al. 2000; Karberg et al. 2001). pACD-GFP-RAM contains a streamlined GFP gene with a CMV promoter and SV40 polyA signal inserted in group II intron domain IV in the orientation opposite intron transcription. The GFP marker is disrupted by a self-splicing *td* group I intron inserted in the forward orientation. During retrotransposition via an RNA intermediate, the *td* intron is spliced, activating the GFP marker, which can be used for screening tadpoles with the intron integrated into a DNA target site. (B) Analysis of *in vitro* RNA splicing products in a denaturing 4% urea polyacrylamide gel. The pACD-GFP-RAM plasmid was linearized with NheI (New England Biolab). "Transcript" represents the transcript from linearized pACD-GFP-RAM vector using a MEGAscript T7 kit (Ambion). The spliced product contains spliced lariat RNA, linear

RNA and small amount of unspliced precursor RNA. All other bands on the gel are intermediates of *td* intron splicing, which occurred during the transcription reaction. (C) RT Real-time PCR analysis of *td* intron splicing efficiency. The spliced RNA (1 mg/ml) was purified and diluted 1,000, 10,000 and 100,000 times to achieve optimal RT PCR amplification efficiency. Two primer sets, $P_{pre}/P_{rev}$ and $P_{sp}/P_{rev}$, were used to detect RNA with the *td* intron (Precursor) and without the *td* intron (Spliced). The blue arrows in panel A show the locations of the primers. The $P_{pre}$ primer specifically anneals to sequences within the *td* intron. The $P_{sp}$ anneals to the ligated junction after splicing of the *td* intron. $P_{rev}$ is the common antisense primer for PCR amplification and also is the primer for the cDNA synthesis in the RT real-time PCR. The y-axis represents the Ct values of the PCR products. The bar graphs show the mean for three determinations. Error bars indicates the standard deviation. The equation used to calculate the *td* splicing efficiency is shown below.

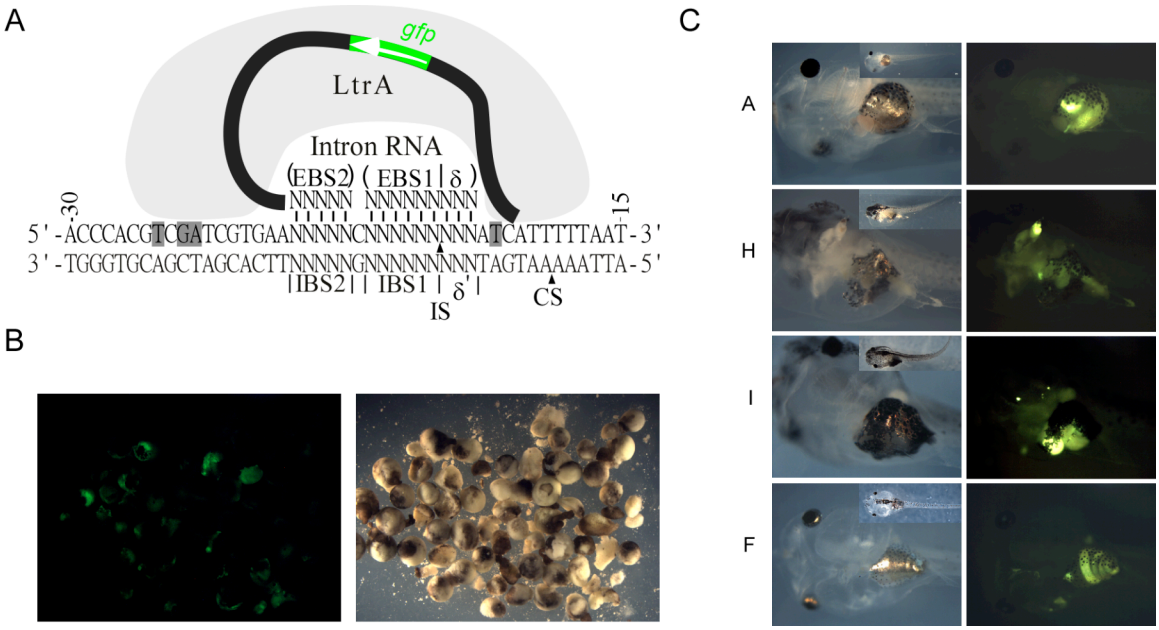



Figure 4.14: Chromosomal integration of the Ll.LtrB-GFP-RAM intron with randomized EBS and δ sequences in *X. laevis*.

(A) Schematic of random GFR-RAM intron library and its interaction with DNA target site. LtrA and the EBS2/1 and δ sequences are involved in DNA target site recognition. The random EBS2/1 and δ sequences in the intron RNA recognize accessible DNA target sites in the genome. As a result, introns insert into different target site within the genome. After intron integration, the GFP gene, which is incorporated into the intron DIV as shown in Figure 4.13, is expressed from its own CMV promoter. Embryos with intron integrations show GFP fluorescence under UV light. The LtrA protein is shown in gray. The black solid line represents linear intron RNA with *gfp* inserted in the opposite orientation. The randomized EBS1, 2 and δ nucleotide residues in the intron RNA indicated as "N". The 45-nt DNA target site is shown with wild-type nucleotide sequences flanking the random IBS1/2 and δ' sequences. The intron insertion site and protein cleavage site are indicated as "IS" and "CS", respectively. (B) Transgenic embryos generated from sperm nuclei treated with the GFP-RAM Ll.LtrB RNP library. The same embryos under UV and bright visible light are shown. (C) Transgenic tadpoles with GFP-RAM intron insertions and the GFP expression in the tadpoles. The same tadpoles are shown under visible and UV light. The whole tadpoles were shown at the right upper corner of visible pictures. The non-uniform expression of GFP presumably reflects the chromosomal location of the integration.

| Condition | Exp. 1 | Exp. 2 | Exp. 3 | Exp. 4 |
|---|---|---|---|---|
| WT Lar | 8.26 | 8.97 | 22.3 | 19.9 |
| WT Lin | | | 0.04 | 0.018 |
| WT GG-Lin | 0.004 | 0.004 | 0.016 | 0.008 |
| WT GG-Lin-2' OMe | 0.002 | 0.005 | | |
| WT GG-Lin-ΔA 2486 | 0.0016 | 0.001 | | 0.002 |
| $DD^-$ GG-Lin | | | | <0.00004 |

Table 4.1: Retrohoming efficiencies of Lar and Lin RNPs in *X. laevis* oocyte nuclei.

Retrohoming assays were done by injecting recipient plasmid pBRR3-ltrB and the indicated Ll.LtrB RNPs into *X. laevis* oocyte nuclei, as described in Fig. 1 and Materials and Methods. Different RNPs contained lariat intron RNA (Lar); linear intron RNA (Lin); linear intron RNA transcribed from a DNA template with two extra 5' G residues (GG-Lin); GG-Lin transcribed from a DNA template with 2'OMe nucleotide residues at the two terminal positions; GG-Lin RNA with the branch-point A-residue deleted (ΔA2486); and GG-Lin RNA + RT-deficient LtrA protein ($DD^-$; YADD→YAAA). After incubating the injected oocytes for 2 h at 25ºC, nucleic acids were transformed into *E. coli* HMS174(DE3) for plating assays. Retrohoming efficiencies were calculated as the ratio of $(Tet^R + Amp^R)/Amp^R$ colonies. These experiments were done in collaboration with Marta Mastroianni.

Exp. 1

| Intron | EBS2 | δ/EBS1 | *X. laevis* | Intron | Occurrence | |
|---|---|---|---|---|---|---|
| 3 | TTTAG | AGG/CAAGCG | ATTTGTTGGTAGTGT | +180 ggtataagttgtgtt | 1 | Protein coding region[b] |
| 4 | TCCTG | TGA/TGTGAT | GTGTAAATTCTCTGG | +1 gtgcgcccagatagg | 1 | Unkown |
| 9 | GCGTA | AGT/GAGCAG | AAGTTGGCGGAGTTG | +1 --gcgcccagatagg | 2 | Protein coding region[a] |
| 12 | NA | NA | CATAGTAGTGTTTTT | +285 gacttatctgttatc | 1 | Protein coding region |
| 13 | TTTAG | AGG/CAAGCG | GCCTCATTCATGTGG | +1 gtgcgcccagatagg | 2 | Unkown |
| 14 | NA | NA | CATAGTAGTGTTTTT | +285 gacttatctgttatc | 1 | Protein coding region |
| 22 | TTCTTC | AGA/TGCTTT | TTAAAGAACAAAGCA | +1 gtgcgcccagatagg | 1 | Unkown |
| 27 | TTGTG | AGA/TGCTTT | TTAAAGAACAAAGCA | +1 gtgcgcccagatagg | 1 | Unkown |
| 28 | CTCCG | CTT/CGCGAG | TCCATCTGTTTTCAA | +1 --gcgcccagatagg | 1 | Unkown |
| 29 | CTGCG | GTT/GCCAAG | CAAAGCTACGGATGG | +67 agccgacctaaccga | 1 | Protein coding region[c] |
| 30 | GACCT | TGC/GGGGGG | GAAAAGCCTACGACA | +1 -----cccagatagg | 1 | 5S rRNA |
| 31 | CGACC | CGA/GTGTGA | TACACAGCATTATGG | +1 gtgcgcccagatagg | 1 | Chloride intracellular channel protein 4 (CLIC4) |

[a] Expressed in anterior neuroectoderm; [b] Similar to endonuclease reverse transcriptase; [c] Expressed in anterior endomedoderm;

Exp. 2

| Intron | EBS2 | δ/EBS1 | *X. laevis* | Intron | Occurrence | |
|---|---|---|---|---|---|---|
| 1 | ATGTG | ATG/GTTGTG | ATACGACTCACTATA | +1 gtgcgcccagatagg | 4 | Cripto-1 (CR1 gene) |
| 6 | TCACG | ACG/TTAACC | GAACGTGACGGTTAA | +1 gtgcgcccagatagg | 1 | Unkown |
| 8 | TTTAG | AAA/GGATCA | GACGCGCTGGGCTAC | +38 gtactactctgtaag | 1 | Unkown |

Exp. 3

| Intron | EBS2 | δ/EBS1 | *X. laevis* | Intron | Tadpole | |
|---|---|---|---|---|---|---|
| 1 | ATGAG | AAG/CTCATG | ACGACTCACTATAGG | +1 gtgcgcccagatagg | F | Unkown |
| 2 | TCACG | ACG/TTAACC | GAACGTGACGGTTAA | +1 gtgcgcccagatagg | H | Unkown |
| 3 | ATGTG | ATG/GTTGTG | ATACGACTCACTATA | +1 gtgcgcccagatagg | A, I | Unkown |

Table 4.2: Summary of integrated Ll.LtrB-GFP-RAM linear introns and its insertion sites in *X. laevis* genome.

The table shows integrated Ll.LtrB-GFP-RAM linear introns in *X. laevis* genome from three independent experiments using GFP-RAM library RNPs. The intron EBS1, 2 and δ sequences and its upstream genomic sequences were identified by TAIL PCR and confirmed by direct PCR. Multiple occurrences of the same intron in different embryos are indicated by "Occurrence". The upstream genomic sequences were blasted against *X. laevis* EST and nucleotide databases in NCBI. Information about the upstream genomic sequences is indicated in the last column. Sequences that did not align with any genes in the above databases are denoted as "Unknown". The intron and *X. laevis* genomic sequences flanking the 5'-integration junctions are indicated in lower and upper case letters respectively. The number in front of the first nucleotide of intron corresponds to its nucleotide position within the intron. In Exp.3, the column "Tadpole" refers to the tadpoles in Figure 4.14C.

# Bibliography


Aizawa, Y., Xiang, Q., Lambowitz, A.M. and Pyle, A.M. 2003. The pathway for DNA recognition and RNA integration by a group II intron retrotransposon. *Mol Cell 11*:795-805.

Amaya, E. and Kroll, K.L. 1999. A method for generating transgenic frog embryos. *Methods Mol Biol 97*:393-414.

Bailey, J.N., Klement, J.F. and McAllister, W.T. 1983. Relationship between promoter structure and template specificities exhibited by the bacteriophage T3 and T7 RNA polymerases. *Proc Natl Acad Sci U S A 80*:2814-2818.

Banks, S., King, S.A., Irvine, D.S. and Saunders, P.T. 2005. Impact of a mild scrotal heat stress on DNA integrity in murine spermatozoa. *Reproduction 129*:505-514.

Banuelos, S., Hierro, A., Arizmendi, J.M., Montoya, G., Prado, A. and Muga, A. 2003. Activation mechanism of the nuclear chaperone nucleoplasmin: role of the core domain. *J Mol Biol 334*:585-593.

Belfort, M., Chandry, P.S. and Pedersen-Lane, J. 1987. Genetic delineation of functional components of the group I intron in the phage T4 td gene. *Cold Spring Harb Symp Quant Biol 52*:181-192.

Belin, P. and Boquet, P.L. 1994. The Escherichia coli dsbA gene is partly transcribed from the promoter of a weakly expressed upstream gene. *Microbiology 140 ( Pt 12)*:3337-3348.

Berthelmann, F. and Bruser, T. 2004. Localization of the Tat translocon components in Escherichia coli. *FEBS Lett 569*:82-88.

Blatny, J.M., Brautaset, T., Winther-Larsen, H.C., Karunakaran, P. and Valla, S. 1997. Improved broad-host-range RK2 vectors useful for high and low regulated gene expression levels in gram-negative bacteria. *Plasmid 38*:35-51.

Blocker, F.J., Mohr, G., Conlan, L.H., Qi, L., Belfort, M. and Lambowitz, A.M. 2005. Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. *RNA 11*:14-28.

Bollag, D.M., Edelstein, S.J. and Rozycki, M.D. 1996. *Protein methods*. New York: Wiley-Liss.

Bronchain, O.J., Hartley, K.O. and Amaya, E. 1999. A gene trap approach in Xenopus. *Curr Biol 9*:1195-1198.

Burland, V., Shao, Y., Perna, N.T., Plunkett, G., Sofia, H.J. and Blattner, F.R. 1998. The complete DNA sequence and analysis of the large virulence plasmid of *Escherichia coli* O157:H7. *Nucleic Acids Res 26*:4196-4204.

Caprara, M.G., Mohr, G. and Lambowitz, A.M. 1996. A tyrosyl-tRNA synthetase protein induces tertiary folding of the group I intron catalytic core. *J Mol Biol 257*:512-531.

Chan, A.W., Luetjens, C.M., Dominko, T., Ramalho-Santos, J., Simerly, C.R., Hewitson, L. and Schatten, G. 2000a. Foreign DNA transmission by ICSI: injection of spermatozoa bound with exogenous DNA results in embryonic GFP expression and live rhesus monkey births. *Mol Hum Reprod 6*:26-33.

Chan, A.W., Luetjens, C.M., Dominko, T., Ramalho-Santos, J., Simerly, C.R., Hewitson, L. and Schatten, G. 2000b. TransgenICSI reviewed: foreign DNA transmission by intracytoplasmic sperm injection in rhesus monkey. *Mol Reprod Dev 56*:325-328.

Chanfreau, G. and Jacquier, A. 1996. An RNA conformational change between the two chemical steps of group II self-splicing. *EMBO J 15*:3466-3476.

Chen, Y., McClane, B.A., Fisher, D.J., Rood, J.I. and Gupta, P. 2005. Construction of an alpha toxin gene knockout mutant of *Clostridium perfringens* type A by use of a mobile group II intron. *Appl Environ Microbiol 71*:7542-7547.

Coros, C.J., Piazza, C.L., Chalamcharla, V.R. and Belfort, M. 2008. A mutant screen reveals RNase E as a silencer of group II intron retromobility in Escherichia coli. *RNA*.

Costa, G.L. and Weiner, M.P. 2003. PCR Primer. In: Dieffenbach, Dveksler, eds. *Cold Spring Harbor Protocols*. NY: Cold Spring Harbor Laboratory Press.

Costa, M., Deme, E., Jacquier, A. and Michel, F. 1997. Multiple tertiary interactions involving domain II of group II self-splicing introns. *J Mol Biol 267*:520-536.

Costa, M., Michel, F. and Westhof, E. 2000. A three-dimensional perspective on exon binding by a group II self-splicing intron. *EMBO J 19*:5007-5018.

Cotten, M., Sealy, L. and Chalkley, R. 1986. Massive phosphorylation distinguishes Xenopus laevis nucleoplasmin isolated from oocytes or unfertilized eggs. *Biochemistry 25*:5063-5069.

Cousineau, B., Smith, D., Lawrence-Cavanagh, S., Mueller, J.E., Yang, J., Mills, D., Manias, D., Dunny, G., Lambowitz, A.M. and Belfort, M. 1998. Retrohoming of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombination. *Cell 94*:451-462.

Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. 2004. WebLogo: a sequence logo generator. *Genome Res 14*:1188-1190.

Cui, X., Matsuura, M., Wang, Q., Ma, H. and Lambowitz, A.M. 2004. A group II intron-encoded maturase functions preferentially in cis and requires both the reverse transcriptase and X domains to promote RNA splicing. *J Mol Biol 340*:211-231.

Dai, L., Toor, N., Olson, R., Keeping, A. and Zimmerly, S. 2003. Database for mobile group II introns. *Nucleic Acids Res 31*:424-426.

Dai, L. and Zimmerly, S. 2002a. Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res 30*:1091-1102.

Dai, L. and Zimmerly, S. 2002b. The dispersal of five group II introns among natural populations of *Escherichia coli*. *RNA 8*:1294-1307.

Daniels, D.L., Michels, W.J., Jr. and Pyle, A.M. 1996. Two competing pathways for self-splicing by group II introns: a quantitative analysis of in vitro reaction rates and products. *J Mol Biol 256*:31-49.

Derijck, A.A., van der Heijden, G.W., Giele, M., Philippens, M.E., van Bavel, C.C. and de Boer, P. 2006. gammaH2AX signalling during sperm chromatin remodelling in the mouse zygote. *DNA Repair (Amst) 5*:959-971.

Draper, G.C. and Gober, J.W. 2002. Bacterial chromosome segregation. *Annu Rev Microbiol 56*:567-597.

Eskes, R., Yang, J., Lambowitz, A.M. and Perlman, P.S. 1997. Mobility of yeast mitochondrial group II introns: engineering a new site specificity and retrohoming via full reverse splicing. *Cell 88*:865-874.

Fedorova, O., Mitros, T. and Pyle, A.M. 2003. Domains 2 and 3 interact to form critical elements of the group II intron active site. *J Mol Biol 330*:197-209.

Frazier, C.L., San Filippo, J., Lambowitz, A.M. and Mills, D.A. 2003. Genetic manipulation of Lactococcus lactis by using targeted group II introns: generation of stable insertions without selection. *Appl Environ Microbiol 69*:1121-1128.

Gaggioli, C., Busca, R., Abbe, P., Ortonne, J.P. and Ballotti, R. 2003. Microphthalmia-associated transcription factor (MITF) is required but is not sufficient to induce the expression of melanogenic genes. *Pigment Cell Res 16*:374-382.

Garrett, J.E. and Carroll, D. 1986. Tx1: a transposable element from Xenopus laevis with some unusual properties. *Mol Cell Biol 6*:933-941.

Gerard, F., Angelini, S. and Wu, L.F. 2002. Export of Thermus thermophilus cytoplasmic beta-glycosidase via the E. coli Tat pathway. *J Mol Microbiol Biotechnol 4*:533-538.

Gorbalenya, A.E. 1994. Self-splicing group I and group II introns encode homologous (putative) DNA endonucleases of a new family. *Protein Sci 3*:1117-1120.

Granlund, M., Michel, F. and Norgren, M. 2001. Mutually exclusive distribution of IS1548 and GBSi1, an active group II intron identified in human isolates of group B streptococci. *J Bacteriol 183*:2560-2569.

Green, M.R. 1986. Pre-mRNA splicing. *Annu Rev Genet 20*:671-708.

Guo, H., Karberg, M., Long, M., Jones, J.P., 3rd, Sullenger, B. and Lambowitz, A.M. 2000. Group II introns designed to insert into therapeutically relevant DNA target sites in human cells. *Science 289*:452-457.

Guo, H., Zimmerly, S., Perlman, P.S. and Lambowitz, A.M. 1997. Group II intron endonucleases use both RNA and protein subunits for recognition of specific sequences in double-stranded DNA. *EMBO J 16*:6835-6848.

Hagmann, M., Bruggmann, R., Xue, L., Georgiev, O., Schaffner, W., Rungger, D., Spaniol, P. and Gerster, T. 1998. Homologous recombination and DNA-end joining reactions in zygotes and early embryos of zebrafish (Danio rerio) and Drosophila melanogaster. *Biol Chem 379*:673-681.

Huang, H.R., Chao, M.Y., Armstrong, B., Wang, Y., Lambowitz, A.M. and Perlman, P.S. 2003. The DIVa maturase binding site in the yeast group II intron aI2 is essential for intron homing but not for in vivo splicing. *Mol Cell Biol 23*:8809-8819.

Ichiyanagi, K., Beauregard, A., Lawrence, S., Smith, D., Cousineau, B. and Belfort, M. 2002. Retrotransposition of the Ll.LtrB group II intron proceeds predominantly via reverse splicing into DNA targets. *Mol Microbiol 46*:1259-1272.

Jacquier, A. and Michel, F. 1990. Base-pairing interactions involving the 5' and 3'-terminal nucleotides of group II self-splicing introns. *J Mol Biol 213*:437-447.

Jacquier, A. and Rosbash, M. 1986. Efficient trans-splicing of a yeast mitochondrial RNA group II intron implicates a strong 5' exon-intron interaction. *Science 234*:1099-1104.

Jarrell, K.A., Dietrich, R.C. and Perlman, P.S. 1988. Group II intron domain 5 facilitates a trans-splicing reaction. *Mol Cell Biol 8*:2361-2366.

Jesuthasan, S. and Subburaju, S. 2002. Gene transfer into zebrafish by sperm nuclear transplantation. *Dev Biol 242*:88-95.

Jiménez-Zurdo, J.I., Garcia-Rodriguez, F.M., Barrientos-Duran, A. and Toro, N. 2003. DNA target site requirements for homing *in vivo* of a bacterial group II intron encoding a protein lacking the DNA endonuclease domain. *J Mol Biol 326*:413-423.

Jones, J.P., 3rd, Kierlin, M.N., Coon, R.G., Perutka, J., Lambowitz, A.M. and Sullenger, B.A. 2005. Retargeting mobile group II introns to repair mutant genes. *Mol Ther 11*:687-694.

Kaneko, T., Moisyadi, S., Suganuma, R., Hohn, B., Yanagimachi, R. and Pelczar, P. 2005. Recombinase-mediated mouse transgenesis by intracytoplasmic sperm injection. *Theriogenology 64*:1704-1715.

Kao, C., Rudisser, S. and Zheng, M. 2001. A simple and efficient method to transcribe RNAs with reduced 3' heterogeneity. *Methods 23*:201-205.

Karberg, M. 2005. Dissertation: Group II Intron Mobility and Its Applications in Biotechnology and Gene Therapy.

Karberg, M., Guo, H., Zhong, J., Coon, R., Perutka, J. and Lambowitz, A.M. 2001. Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. *Nat Biotechnol 19*:1162-1167.

Kennell, J.C., Moran, J.V., Perlman, P.S., Butow, R.A. and Lambowitz, A.M. 1993. Reverse transcriptase activity associated with maturase-encoding group II introns in yeast mitochondria. *Cell 73*:133-146.

Koch, J.L., Boulanger, S.C., Dib-Hajj, S.D., Hebbar, S.K. and Perlman, P.S. 1992. Group II introns deleted for multiple substructures retain self-splicing activity. *Mol Cell Biol 12*:1950-1958.

Koga, A., Inagaki, H., Bessho, Y. and Hori, H. 1995. Insertion of a novel transposable element in the tyrosinase gene is responsible for an albino mutation in the medaka fish, Oryzias latipes. *Mol Gen Genet 249*:400-405.

Kroll, K.L. and Amaya, E. 1996. Transgenic Xenopus embryos from sperm nuclear transplantations reveal FGF signaling requirements during gastrulation. *Development 122*:3173-3183.

Kumasaka, M., Sato, H., Sato, S., Yajima, I. and Yamamoto, H. 2004. Isolation and developmental expression of Mitf in Xenopus laevis. *Dev Dyn 230*:107-113.

Kumasaka, M., Sato, S., Yajima, I., Goding, C.R. and Yamamoto, H. 2005. Regulation of melanoblast and retinal pigment epithelium development by Xenopus laevis Mitf. *Dev Dyn 234*:523-534.

Kumasaka, M., Sato, S., Yajima, I. and Yamamoto, H. 2003. Isolation and developmental expression of tyrosinase family genes in Xenopus laevis. *Pigment Cell Res 16*:455-462.

Lambowitz, A.M., Mohr, G. and Zimmerly, S. 2005. Group II Intron Homing Endonucleases: Ribonucleoprotein Complexes with Programmable Target Specificity. *Homing Endonucleases and Inteins*: Springer-Verlag Berlin Heidelberg. pp 121-145.

Lambowitz, A.M. and Zimmerly, S. 2004. Mobile group II introns. *Annu Rev Genet 38*:1-35.

Lee, P.A., Tullman-Ercek, D. and Georgiou, G. 2006. The bacterial twin-arginine translocation pathway. *Annu Rev Microbiol 60*:373-395.

Lehmann, K. and Schmidt, U. 2003. Group II introns: structure and catalytic versatility of large natural ribozymes. *Crit Rev Biochem Mol Biol 38*:249-303.

Leno, G.H., Mills, A.D., Philpott, A. and Laskey, R.A. 1996. Hyperphosphorylation of nucleoplasmin facilitates Xenopus sperm decondensation at fertilization. *J Biol Chem 271*:7253-7256.

Lister, J.A., Robertson, C.P., Lepage, T., Johnson, S.L. and Raible, D.W. 1999. nacre encodes a zebrafish microphthalmia-related protein that regulates neural-crest-derived pigment cell fate. *Development 126*:3757-3767.

Liu, Y.G., Mitsukawa, N., Oosumi, T. and Whittier, R.F. 1995. Efficient isolation and mapping of Arabidopsis thaliana T-DNA insert junctions by thermal asymmetric interlaced PCR. *Plant J 8*:457-463.

Majumder, H.K., Maitra, U. and Rosenberg, M. 1979. Termination of transcription by bacteriophage T3 RNA polymerase: homogeneous 3'-terminal oligonucleotide sequence of in vitro T3 RNA polymerase transcripts. *Proc Natl Acad Sci U S A 76*:5110-5113.

Malhotra, M. and Srivastava, S. 2008. An ipdC gene knock-out of Azospirillum brasilense strain SM and its implications on indole-3-acetic acid biosynthesis and plant growth promotion. *Antonie Van Leeuwenhoek 93*:425-433.

Malik, H.S., Burke, W.D. and Eickbush, T.H. 1999. The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol 16*:793-805.

Martinez-Abarca, F., Barrientos-Duran, A., Fernandez-Lopez, M. and Toro, N. 2004. The RmInt1 group II intron has two different retrohoming pathways for mobility using predominantly the nascent lagging strand at DNA replication forks for priming. *Nucleic Acids Res 32*:2880-2888.

Martinez-Abarca, F. and Toro, N. 2000. RecA-independent ectopic transposition in vivo of a bacterial group II intron. *Nucleic Acids Res 28*:4397-4402.

Mastroianni, M., Watanabe, K., White, T.B., Zhuang, F., Vernon, J., Matsuura, M., Wallingford, J. and Lambowitz, A.M. 2008. Group II Intron-Based Gene Targeting Reactions in Eukaryotes. *PloS ONE 3*:e3121.

Matsuura, M., Noah, J.W. and Lambowitz, A.M. 2001. Mechanism of maturase-promoted group II intron splicing. *EMBO J 20*:7259-7270.

Matsuura, M., Saldanha, R., Ma, H., Wank, H., Yang, J., Mohr, G., Cavanagh, S., Dunny, G.M., Belfort, M. and Lambowitz, A.M. 1997. A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities:

biochemical demonstration of maturase activity and insertion of new genetic information within the intron. *Genes Dev 11*:2910-2924.

Michel, F. and Ferat, J.L. 1995. Structure and activities of group II introns. *Annu Rev Biochem 64*:435-461.

Michel, F., Umesono, K. and Ozeki, H. 1989. Comparative and functional anatomy of group II catalytic introns--a review. *Gene 82*:5-30.

Milligan, J.F., Groebe, D.R., Witherell, G.W. and Uhlenbeck, O.C. 1987. Oligoribonucleotide synthesis using T7 RNA polymerase and synthetic DNA templates. *Nucleic Acids Res 15*:8783-8798.

Mills, D.A., Manias, D.A., McKay, L.L. and Dunny, G.M. 1997. Homing of a group II intron from Lactococcus lactis subsp. lactis ML3. *J Bacteriol 179*:6107-6111.

Mills, D.A., McKay, L.L. and Dunny, G.M. 1996. Splicing of a group II intron involved in the conjugative transfer of pRS01 in lactococci. *J Bacteriol 178*:3531-3538.

Mohr, G., Perlman, P.S. and Lambowitz, A.M. 1993. Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucleic Acids Res 21*:4991-4997.

Mohr, G., Smith, D., Belfort, M. and Lambowitz, A.M. 2000. Rules for DNA target-site recognition by a lactococcal group II intron enable retargeting of the intron to specific DNA sequences. *Genes Dev 14*:559-573.

Mohr, G., Zhang, A., Gianelos, J.A., Belfort, M. and Lambowitz, A.M. 1992. The neurospora CYT-18 protein suppresses defects in the phage T4 td intron by stabilizing the catalytically active structure of the intron core. *Cell 69*:483-494.

Morl, M. and Schmelzer, C. 1990. Integration of group II intron bI1 into a foreign RNA by reversal of the self-splicing reaction in vitro. *Cell 60*:629-636.

Murray, R.G., Steed, P. and Elson, H.E. 1965. The Location of the Mucopeptide in Sections of the Cell Wall of Escherichia Coli and Other Gram-Negative Bacteria. *Can J Microbiol 11*:547-560.

Nieuwkoop, P.D.a.F., J. 1994. *Normal Table of Xenopus laevis (Daudin)*. New York: Garland Publishing Inc.

Nikaido, H. 1994. Isolation of outer membranes. *Methods Enzymol 235*:225-234.

Niki, H. and Hiraga, S. 1998. Polar localization of the replication origin and terminus in Escherichia coli nucleoids during chromosome partitioning. *Genes Dev 12*:1036-1045.

Noah, J.W., Park, S., Whitt, J.T., Perutka, J., Frey, W. and Lambowitz, A.M. 2006. Atomic force microscopy reveals DNA bending during group II intron

ribonucleoprotein particle integration into double-stranded DNA. *Biochemistry 45*:12424-12435.

Oetting, W.S. and King, R.A. 1993. Molecular basis of type I (tyrosinase-related) oculocutaneous albinism: mutations and polymorphisms of the human tyrosinase gene. *Hum Mutat 2*:1-6.

Osborn, M.J., Gander, J.E. and Parisi, E. 1972a. Mechanism of assembly of the outer membrane of Salmonella typhimurium. Site of synthesis of lipopolysaccharide. *J Biol Chem 247*:3973-3986.

Osborn, M.J., Gander, J.E., Parisi, E. and Carson, J. 1972b. Mechinism of assembly of the outer membrane of Salmonella typhimurium. Isolation and characterization of cytoplasmic and outer membrane. *J Biol Chem 247*:3962-3972.

Osborn, M.J. and Munson, R. 1974. Separation of the inner (cytoplasmic) and outer membranes of Gram-negative bacteria. *Methods Enzymol 31*:642-653.

Paul, C., Murray, A., Spears, N. and Saunders, P. 2008. A single, mild, transient scrotal heat stress causes DNA damage, subfertility and impairs formation of blastocysts in mice. *Reproduction*.

Pearson, M.M. and Mobley, H.L. 2007. The type III secretion system of Proteus mirabilis HI4320 does not contribute to virulence in the mouse model of ascending urinary tract infection. *J Med Microbiol 56*:1277-1283.

Perlman, P.S. and Podar, M. 1996. Reactions catalyzed by group II introns in vitro. *Methods Enzymol 264*:66-86.

Perry, A.C., Wakayama, T., Kishikawa, H., Kasai, T., Okabe, M., Toyoda, Y. and Yanagimachi, R. 1999. Mammalian transgenesis by intracytoplasmic sperm injection. *Science 284*:1180-1183.

Perutka, J., Wang, W., Goerlitz, D. and Lambowitz, A.M. 2004. Use of computer-designed group II introns to disrupt *Escherichia coli* DExH/D-box protein and DNA helicase genes. *J Mol Biol 336*:421-439.

Plante, I. and Cousineau, B. 2006. Restriction for gene insertion within the Lactococcus lactis Ll.LtrB group II intron. *RNA 12*:1980-1992.

Podar, M., Chu, V.T., Pyle, A.M. and Perlman, P.S. 1998. Group II intron splicing in vivo by first-step hydrolysis. *Nature 391*:915-918.

Porteus, M.H. and Carroll, D. 2005. Gene targeting using zinc finger nucleases. *Nat Biotechnol 23*:967-973.

Pyle, A. and Lambowitz, A.M. 2006. Group II introns: ribozymes that splice RNA and invade DNA. In: Gesteland RF, Cech TR, Atkins JF, eds. *The RNA world*. Cold Spring Harbor: Cold Spring Harbor Laboratory Press. pp 469-505.

Rieth, A., Pothier, F. and Sirard, M.A. 2000. Electroporation of bovine spermatozoa to carry DNA containing highly repetitive sequences into oocytes and detection of homologous recombination events. *Mol Reprod Dev 57*:338-345.

Rinchik, E.M., Stoye, J.P., Frankel, W.N., Coffin, J., Kwon, B.S. and Russell, L.B. 1993. Molecular analysis of viable spontaneous and radiation-induced albino (c)-locus mutations in the mouse. *Mutat Res 286*:199-207.

Rodriguez, S.A., Yu, J.J., Davis, G., Arulanandam, B.P. and Klose, K.E. 2008. Targeted inactivation of francisella tularensis genes by group II introns. *Appl Environ Microbiol 74*:2619-2626.

Roger, B., Moisand, A., Amalric, F. and Bouvet, P. 2002. rDNA transcription during Xenopus laevis oogenesis. *Biochem Biophys Res Commun 290*:1151-1160.

Roitzsch, M. and Pyle, A.M. 2009. The linear form of a group II intron catalyzes efficient autocatalytic reverse splicing, establishing a potential for mobility. *RNA 15*:473-482.

Saldanha, R., Chen, B., Wank, H., Matsuura, M., Edwards, J. and Lambowitz, A.M. 1999. RNA and protein catalysis in group II intron splicing and mobility reactions using purified components. *Biochemistry 38*:9069-9083.

San Filippo, J. and Lambowitz, A.M. 2002. Characterization of the C-terminal DNA-binding/DNA endonuclease region of a group II intron-encoded protein. *J Mol Biol 324*:933-951.

Schmelzer, C. and Schweyen, R.J. 1986. Self-splicing of group II introns in vitro: mapping of the branch point and mutational inhibition of lariat formation. *Cell 46*:557-565.

Shao, L., Hu, S., Yang, Y., Gu, Y., Chen, J., Jiang, W. and Yang, S. 2007. Targeted gene disruption by use of a group II intron (targetron) vector in Clostridium acetobutylicum. *Cell Res 17*:963-965.

Shearman, C., Godon, J.J. and Gasson, M. 1996. Splicing of a group II intron in a functional transfer gene of Lactococcus lactis. *Mol Microbiol 21*:45-53.

Simon, D.M., Clarke, N.A., McNeil, B.A., Johnson, I., Pantuso, D., Dai, L., Chai, D. and Zimmerly, S. 2008. Group II introns in eubacteria and archaea: ORF-less introns and new varieties. *RNA 14*:1704-1713.

Singh, N.N. and Lambowitz, A.M. 2001. Interaction of a group II intron ribonucleoprotein endonuclease with its DNA target site investigated by DNA footprinting and modification interference. *J Mol Biol 309*:361-386.

Sive, H.L., Grainger, R.M. and Harland, R.M. 2000. *Early development of Xenopus laevis : a laboratory manual*. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

Smith, D., Zhong, J., Matsuura, M., Lambowitz, A.M. and Belfort, M. 2005. Recruitment of host functions suggests a repair pathway for late steps in group II intron retrohoming. *Genes Dev 19*:2477-2487.

Smith, K. and Spadafora, C. 2005. Sperm-mediated gene transfer: applications and implications. *Bioessays 27*:551-562.

Smith, S.J., Fairclough, L., Latinkic, B.V., Sparrow, D.B. and Mohun, T.J. 2006. Xenopus laevis transgenesis by sperm nuclear injection. *Nat Protoc 1*:2195-2203.

Sparrow, D.B., Latinkic, B. and Mohun, T.J. 2000. A simplified method of generating transgenic Xenopus. *Nucleic Acids Res 28*:E12.

Strandberg, L. and Enfors, S.O. 1991. Factors influencing inclusion body formation in the production of a fused protein in Escherichia coli. *Appl Environ Microbiol 57*:1669-1674.

Sugimoto, N., Nakano, M. and Nakano, S. 2000. Thermodynamics-structure relationship of single mismatches in RNA/DNA duplexes. *Biochemistry 39*:11270-11281.

Tabak, H.F., Van der Horst, G., Winter, A.J., Smit, J., Van der Veen, R., Kwakman, J.H., Grivell, L.A. and Arnberg, A.C. 1987. Reactions mediated by yeast mitochondrial group I and II introns. *Cold Spring Harb Symp Quant Biol 52*:213-221.

Toor, N., Hausner, G. and Zimmerly, S. 2001. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA 7*:1142-1152.

Toor, N., Keating, K.S., Taylor, S.D. and Pyle, A.M. 2008. Crystal structure of a self-spliced group II intron. *Science 320*:77-82.

Toro, N. 2003. Bacteria and Archaea Group II introns: additional mobile genetic elements in the environment. *Environ Microbiol 5*:143-151.

Toro, N., Jimenez-Zurdo, J.I. and Garcia-Rodriguez, F.M. 2007. Bacterial group II introns: not just splicing. *FEMS Microbiol Rev 31*:342-358.

Vachtenheim, J. and Drdova, B. 2004. A dominant negative mutant of microphthalmia transcription factor (MITF) lacking two transactivation domains suppresses transcription mediated by wild type MITF and a hyperactive MITF derivative. *Pigment Cell Res 17*:43-50.

van der Veen, R., Kwakman, J.H. and Grivell, L.A. 1987. Mutations at the lariat acceptor site allow self-splicing of a group II intron without lariat formation. *EMBO J 6*:3827-3831.

Vogel, J. and Borner, T. 2002. Lariat formation and a hydrolytic pathway in plant chloroplast group II intron splicing. *EMBO J 21*:3794-3803.

Wank, H., SanFilippo, J., Singh, R.N., Matsuura, M. and Lambowitz, A.M. 1999. A reverse transcriptase/maturase promotes splicing by binding at its own coding segment in a group II intron RNA. *Mol Cell 4*:239-250.

Weterings, E. and van Gent, D.C. 2004. The mechanism of non-homologous end-joining: a synopsis of synapsis. *DNA Repair (Amst) 3*:1425-1435.

Xiang, Q., Qin, P.Z., Michels, W.J., Freeland, K. and Pyle, A.M. 1998. Sequence specificity of a group II intron ribozyme: multiple mechanisms for promoting unusually high discrimination against mismatched targets. *Biochemistry 37*:3839-3849.

Yang, J., Mohr, G., Perlman, P.S. and Lambowitz, A.M. 1998. Group II intron mobility in yeast mitochondria: target DNA-primed reverse transcription activity of aI1 and reverse splicing into DNA transposition sites *in vitro*. *J Mol Biol 282*:505-523.

Yao, J. and Lambowitz, A.M. 2007. Gene targeting in gram-negative bacteria by use of a mobile group II intron ("Targetron") expressed from a broad-host-range vector. *Appl Environ Microbiol 73*:2735-2743.

Yao, J., Zhong, J., Fang, Y., Geisinger, E., Novick, R.P. and Lambowitz, A.M. 2006. Use of targetrons to disrupt essential and nonessential genes in Staphylococcus aureus reveals temperature sensitivity of Ll.LtrB group II intron splicing. *RNA 12*:1271-1281.

Yao, J., Zhong, J. and Lambowitz, A.M. 2005. Gene targeting using randomly inserted group II introns (targetrons) recovered from an Escherichia coli gene disruption library. *Nucleic Acids Res 33*:3351-3362.

Zernicka-Goetz, M., Pines, J., Ryan, K., Siemering, K.R., Haseloff, J., Evans, M.J. and Gurdon, J.B. 1996. An indelible lineage marker for Xenopus using a mutated green fluorescent protein. *Development 122*:3719-3724.

Zhao, J. and Lambowitz, A.M. 2005. Inaugural Article: A bacterial group II intron-encoded reverse transcriptase localizes to cellular poles. *Proc Natl Acad Sci U S A 102*:16133-16140.

Zhao, J., Niu, W., Yao, J., Mohr, S., Marcotte, E.M. and Lambowitz, A.M. 2008. Group II intron protein localization and insertion sites are affected by polyphosphate. *PLoS Biol 6*:e150.

Zhong, J., Karberg, M. and Lambowitz, A.M. 2003. Targeted and random bacterial gene disruption using a group II intron (targetron) vector containing a retrotransposition-activated selectable marker. *Nucleic Acids Res 31*:1656-1664.

Zhong, J. and Lambowitz, A.M. 2003. Group II intron mobility using nascent strands at DNA replication forks to prime reverse transcription. *EMBO J 22*:4555-4565.

Zhong, J. and Lambowitz, A.M. 2003b. Group II intron mobility using nascent strands at DNA replication forks to prime reverse transcription. *EMBO J 22*:4555-4565.

Zhuang, F., Karberg, M., Perutka, J. and Lambowitz, A.M. 2009. EcI5, a group IIB intron with high retrohoming frequency: DNA target site recognition and use in gene targeting. *RNA 15*:432-449.

Zimmerly, S., Guo, H., Eskes, R., Yang, J., Perlman, P.S. and Lambowitz, A.M. 1995a. A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility. *Cell 83*:529-538.

Zimmerly, S., Hausner, G. and Wu, X. 2001. Phylogenetic relationships among group II intron ORFs. *Nucleic Acids Res 29*:1238-1250.

Zimmerly, S., Moran, J.V., Perlman, P.S. and Lambowitz, A.M. 1999. Group II intron reverse transcriptase in yeast mitochondria. Stabilization and regulation of reverse transcriptase activity by the intron RNA. *J Mol Biol 289*:473-490.

# Vita

Fanglei Zhuang was born in Ningbo, Zhejiang province, China, on May 20th 1980. She is the elder daughter of Anlin Zhuang and Liying Huang. After her graduating from Xiaoshi high school, Ningbo, in 1998, she was admitted by Zhejiang University, Hangzhou, where she received the degree of Bachelor of Science in Biology in June 2002. Two months after her graduation, she started her graduate school in Microbiology department of the University of Texas at Austin, Austin, USA. In August 2009, she received her degree Doctor of Philosophy from the University of Texas at Austin. She coauthored the following publications:

Liang, W., Mao, B., Li, D., Li, S. and **Zhuang, F.** 2001. Activity Change and Distribution of β-cyanoalanine Synthase in Rice Seeds during Germination. *Acta Phytophysiologica Sinica* 27(3): 243-247.

Mastroianni, M., Watanabe, K., White, T.B., **Zhuang, F.**, Vernon, J., Matsuura, M., Wallingford, J. and Lambowitz, A.M. 2008. Group II Intron-Based Gene Targeting Reactions in Eukaryotes. *PloS ONE* 3:e3121.

**Zhuang, F.**, Watanabe, K., Mastroianni, M., Wallingford, J. and Lambowitz, A.M. 2008. Abstract for 2008 Transgenic Technology Meeting: Gene targeting in *Xenopus laevis* via site-specific modification of sperm DNA using mobile group II introns ("targetrons"). *Transgenic Research* 17: 1023.

**Zhuang, F.**, Karberg, M., Perutka, J. and Lambowitz, A.M. 2009. EcI5, a group IIB intron with high retrohoming frequency: DNA target site recognition and use in gene targeting. *RNA 15*:432-449.

**Zhuang, F.***, Mastroianni, M.*, White, T.B. and Lambowitz, A.M. 2009. Linear group II introns RNAs can retrohome in eukaryotes by a mechanism involving non-homologous end-joining for cDNA ligation. *In preparation.*

Permanent address: Tai Ping Qiao #20, Ci Cheng Zeng, Jiang Bei Qu, Ningbo City, Zhejiang Province, China, 315031.

This dissertation was typed by Fanglei Zhuang.